BANK OF CANADA
BANQUE DU CANADA

**Staff Working Paper/Document de travail du personnel — 2021-37**

Last updated: August 5, 2021

# Estimating Large-Dimensional Connectedness Tables: The Great Moderation Through the Lens of Sectoral Spillovers

by Felix Brunner[1] and Ruben Hipp[2]

1 Nova School of Business and Economics

2 Financial Stability Department
Bank of Canada, Ottawa, Ontario, Canada K1A 0G9

felix.brunner@novasbe.pt, rhipp@bankofcanada.ca



ISSN 1701-9397

# Acknowledgements

We are thankful for comments from Jason Allen, Tatjana Dahlhaus, Thibaut Duprey, Soojin Jo, Paulo M. M. Rodrigues, and various seminar participants. This work was supported by the Fundação para a Ciência e Tecnologia, Portugal. A special thanks goes to Matteo Barigozzi and Christian Brownlees, without whom this work would not exist.

# Abstract

We estimate sectoral spillovers around the Great Moderation with the help of forecast error variance decomposition tables. Obtaining such tables in high dimensions is challenging since they are functions of the estimated vector autoregressive coefficients and the residual covariance matrix. In a simulation study, we compare various regularization methods for both and conduct a comprehensive analysis of their performance. We show that standard estimators of large connectedness tables lead to biased results and high estimation uncertainty, which can both be mitigated by regularization. To explore possible causes for the Great Moderation, we apply a cross-validated estimator on sectoral spillovers of industrial production in the US from 1972 to 2007. We find that a handful of sectors considerably decreased their outgoing links, which hints at a complimentary explanation for the Great Moderation.



*Topics: Business fluctuations and cycles, Econometric and statistical methods*
*JEL codes*: *C22, C52, E23, E27*

# 1 Introduction

With the onset of the Great Moderation, around 1984, key macroeconomic time series exhibit sharp decreases in growth rate volatility. Whether this shift in fluctuations is due to a structural change in the economy, improved economic policies, or just good luck has been extensively studied using a variety of approaches. In particular, for industrial production (IP), the literature provides manifold narratives, often using contemporaneous correlations between sectors to approximate dependencies. Recent advancements further allow econometricians to describe directional dependencies in the form of forecast error variance decompositions (FEVDs) (see Diebold and Yılmaz, 2014). Namely, FEVDs measure how much the variation of one sector can explain the variation of another. They do so by condensing contemporaneous and lagged dependencies into a single connectedness table based on vector autoregressions (VARs). These tables, however, inherit the VAR's estimation uncertainty in high dimensions: the estimation error blows up when the number of variables ($N$) approaches the number of time observations ($T$).

The goal of this paper is to compare regularization methods for estimating large networks in a time series context. It consists of two parts: an extensive Monte Carlo (MC) simulation and an application to sectoral spillovers of IP.

Econometricians typically face a high-dimensional setup when estimating FEVDs. This is going to be the case in our application to IP spillovers. For one thing, networks must include all relevant variables to have a unified and precise interpretation; i.e., $N$ is large. For another thing, time variation in the parameters — e.g., a structural break or rolling windows — reduces the number of observations; i.e., $T$ is often small. With large $N$ and small $T$, standard estimation methods produce poor estimates and bad forecasts due to overfitting. Regularization methods for regressions and covariance matrices counteract these adverse effects. Yet, it remains unclear which ones to choose. Thus, we provide an extensive MC study of regularization techniques combined with FEVDs to guide researchers in large-dimensional network estimations.

While there are successful applications of FEVDs in high dimensions (e.g., see Demirer et al., 2017), two questions remain unanswered. First, how does estimation uncertainty affect the overall results of FEVDs? Second, would an additional regularization of the innovation covariance matrix improve the results? Our MC simulation results demonstrate the performance gain of each regularization step regarding the estimation of the overall FEVD network. While regularization mostly boils down to improving the out-of-sample fit by trading unbiasedness for a reduction in variance, we show that for the case of FEVDs, regularization methods also decrease the estimation bias through improved forecasts. More precisely, regularization of both the coefficients and the covariance matrix not only reduces the estimation uncertainty of the FEVD but also mitigates a positive bias in the entries. To the best of our knowledge, this result is novel in the literature

and highlights the importance of regularization in the context of FEVDs. Perhaps surprisingly, a horse race of regularization methods applied to FEVDs yields similar results among estimators, with the winner being conditional on the data generating process.

In the application, we investigate changes to the dependency structure of IP across sectors around the Great Moderation. The central question is whether the large volatility of the aggregate IP index originated from amplifications of sectoral fluctuations by the asymmetric network, in the spirit of Acemoglu et al. (2012). That is, we address the gap in the literature on how structural change in sectoral interconnectedness affects aggregate index volatility. The directed nature of FEVDs allows us to investigate asymmetric linkages that were previously hidden under the rationale of correlations as common exposure to aggregate shocks. Thus, we motivate this application by the fact that a strong intersectoral network of idiosyncratic shocks is observationally equivalent to the prevalence of aggregate shocks, if connectedness is ignored.

We apply FEVDs to scrutinize IP spillovers between 88 sectors in the US from 1972 to 2007. The estimation is challenging since the split into pre- and post-Great Moderation periods reduces the effective sample size. We use cross-validated regularization to tackle this challenge. The estimates provide time-varying spillover networks that uncover the corresponding asymmetric dependency structure among sectors. We find that sectors that were influential initially decreased their outgoing links after 1984, reducing the importance of spillover effects overall. An analysis of contributions to aggregate variation suggests that a handful of sectors considerably added to the high volatility pre-Great Moderation, supporting a narrative of structural change that is complimentary to the story of aggregate shock volatility.

Our comparative MC study connects to the literature on the regularization of regressions and covariance matrices. Regularizations have become popular not only because the increasing availability of data makes variable selection more critical but also because technological advances make the application of high-dimensional estimators increasingly feasible. For the regression step, we consider shrinkage estimators, such as the ridge regression by Hoerl and Kennard (1970), variable selection methods, such as the LASSO by Tibshirani (1996), and combinations of the two, such as the adaptive elastic-net by Zou and Zhang (2009). Whereas these methods mostly find application in cross-sectional contexts, the time series literature has succeeded in using regularization, for example, in the general case of VARs in Kascha and Trenkler (2015) and in the setting of FEVDs in Demirer et al. (2017). For the regularization of covariance matrices, we examine variable selection methods for the partial correlation matrix, as in the graphical LASSO by Friedman et al. (2008), optimal shrinkage estimators, as in Ledoit and Wolf (2004), and sample covariance thresholding, as in Bickel and Levina (2008), Rothman et al. (2009) and Cai and Liu (2011). We contribute to this diverse literature by comparing their respective performances in the context of FEVDs.

Additionally, our empirical results contribute to the understanding of the changes that took place during the Great Moderation. By highlighting the transition from a centralized economy that heavily relies on the productivity of a few sectors to a more diversified network, we offer a unifying view on opposing explanations of the decline in production volatility. Namely, Foerster et al. (2011) argues that this phenomenon is explained by a decline in aggregate shocks; i.e., shocks to multiple sectors at once. This contrasts with the results of Gabaix (2011), who considers overweight index constituents as a central driver of aggregate fluctuations. Our empirical results show that the intersectoral network has changed structurally with the Great Moderation, making some idiosyncratic shocks less likely to impact the aggregate. In other words, before the Great Moderation, granular shocks were able to propagate through the network so far that they amplified to strong aggregate fluctuations. Our description of the dynamics of the network structure therefore contributes to the debate over the origins of the strong aggregate fluctuations that occurred before the Great Moderation.

The rest of the paper is organized as follows. In Section 2, we introduce the concepts of FEVDs and provide an overview of various regularization methods. We assess their performance with a simulation study in Section 3. Section 4 applies the regularization of FEVDs to the IP setup to answer the question of sectoral spillovers. Finally, Section 5 concludes.

# 2 Methodology

This section provides a general overview of FEVDs and introduces suitable regularization methodologies to mitigate estimation uncertainty in large-dimensional applications.

We start with an $N$-dimensional stable VAR(1) process,

$$y_t = \nu + Ay_{t-1} + u_t, \quad \forall t = 1, ..., T, \tag{1}$$

where $u_t$ has multivariate normal distribution with covariance matrix $\Sigma$ and $\nu$ is a fixed vector of intercept terms. Following Pesaran and Shin (1998), we obtain the FEVD as a function of the VAR coefficient matrix $A$ and the innovation covariance matrix $\Sigma$.[1] A detailed derivation is presented in Section 2.1.

As in many structurally motivated economic models, we want to include a broad set of variables. However, considering many variables leads to a curse of dimensionality. Thus, to estimate high-dimensional FEVDs, we have to estimate the VAR coefficient matrix and the innovation covariance matrix in (1) in large dimensions. For that purpose, we assess the performance of various estimation techniques by comparing different regularization

[1]Note that any VAR(p) translates into a VAR(1), hence, the companion form generalizes to higher-order lag numbers. For details see Lütkepohl (2005, page 15).

approaches for large VARs.[2] Figure 1 gives an overview of the regularization methods considered in this paper.

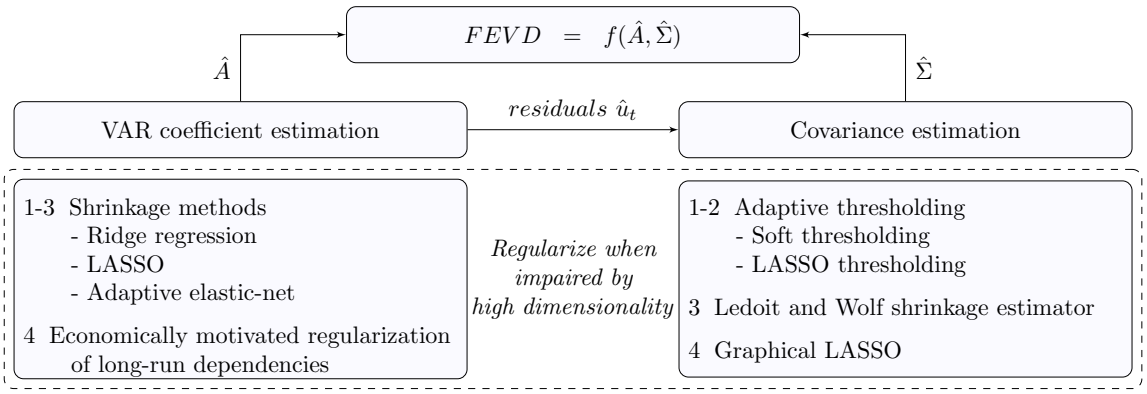


Figure 1: Estimation overview for high-dimensional forecast error variance decompositions (FEVDs). The estimates $\hat{A}$ and $\hat{\Sigma}$ are required for the estimation of FEVDs. The lists describe the considered regularization methods in this paper.

The statistical learning literature contains two types of regularization techniques that are beneficial for FEVDs: the regularization of coefficients and the regularization of covariance matrices. Concerning the first, LASSO techniques tend to perform well in autoregressive setups. Such techniques apply regularization to the coefficient matrix $A$ but do not imply any regularization for the covariance matrix $\Sigma$. The combination of $\hat{A}$ and $\hat{\Sigma}$ in the FEVD, however, suggests that poor estimation of the innovation covariance matrix renders the overall estimate noisy. That is, the estimation of $\Sigma$ suffers from the same uncertainty induced by high dimensionality. Hence, we resort to covariance shrinkage estimators as an alternative to the sample covariance matrix. Our approach, therefore, is to combine regularization methods for the unknown regression coefficient and covariance matrix to achieve the best possible estimate of the FEVD with respect to the estimation error. Section 2.2 describes each of the methods in detail.

[2]We address the most prominent examples, but we are aware that there are additional approaches that are beyond the scope of this paper. L0 penalties in high-dimensional regressions are computationally infeasible for classical machines: the non-convexity of the L0 penalty creates a combinatorial problem, which is NP-hard. Notable examples of variable selection are the information criteria AIC/BIC and the Variational Garotte in Kappen and Gómez (2014). We experimented with the latter approach in small scale simulations, which turned out to be computationally too expensive and did not yield any improvement. The covariance matrix estimation literature also deals with the regularization of the eigenvalues; e.g., in Lam et al. (2016).

## 2.1 Generalized Forecast Error Variance Decompositions

Due to the stability assumption, the VAR process in (1) can be written in moving average (MA) representation

$$y_t = \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \tag{2}$$

where the MA parameters are defined as $\Phi_i = A^i$, $\forall i \geq 0$ and $\mu$ represents the mean term. Note that the components of $u_t$ are generally not orthogonal such that structural interpretations are economically meaningless. Koop et al. (1996) and Pesaran and Shin (1998) define the (unscaled) generalized impulse response (IR) function at horizon $h$ to an impulse $\Delta_j$ on the $j$th entry of the reduced form innovation $u_t$. They do so by integrating out the effect of all of the remaining shocks in the shock vector:

$$\mathsf{IR}(h, \Delta_j, j) = E[y_{t+h}|u_{j,t} = \Delta_j] - E[y_{t+h}|u_{j,t} = 0]. \tag{3}$$

Under a Gaussian assumption, we can use

$$E[u_t|u_{j,t} = \Delta_j] = (\sigma_{1j}, ..., \sigma_{Nj})\sigma_{jj}^{-1}\delta_j = \Sigma e_j \sigma_{jj}^{-1}\Delta_j,$$

with $\sigma_{ij}$ being the $ij$th entry in $\Sigma$ and $e_j$ as the $j$th column of the identity matrix. Substituting this expression in (3), the impulse response function can be rewritten as

$$\mathsf{IR}(h, \Delta_j, j) = \Phi_h \Sigma e_j \sigma_{jj}^{-1}\Delta_j.$$

It is customary to set $\Delta_j = \sqrt{\sigma_{jj}}$, which yields the scaled generalized impulse response functions $\mathsf{IR}(h, \sqrt{\sigma_{jj}}, j)$. We assemble the scaled IRs in the $(N \times N)$ matrix

$$\Psi^g(h) = [\psi_{ij}^g(h)] = [\mathsf{IR}(h, \sqrt{\sigma_{11}}, 1), ..., \mathsf{IR}(h, \sqrt{\sigma_{NN}}, N)] = \Phi_h \Sigma \, \mathrm{diag}(\Sigma)^{-\frac{1}{2}}, \tag{4}$$

where $\mathrm{diag}(M)$ denotes a diagonal matrix with diagonal values of any square matrix $M$.

Analog to standard impulse response analysis, we gain further insights rewriting $y_t$ as a vector-valued impulse response function multiplied by an innovation vector. Let $\mathrm{P} = \Sigma \, \mathrm{diag}(\Sigma)^{-\frac{1}{2}}$ and define the generalized shock $u_t^g$ as

$$u_t^g := \mathrm{P}^{-1} u_t \sim \mathcal{N}\left(\mathbf{0}_{N\times 1}, \Omega\right),$$

where $\Omega = \mathrm{diag}(\Sigma)^{\frac{1}{2}}\Sigma^{-1}\mathrm{diag}(\Sigma)^{\frac{1}{2}}$. Then it is possible to express (2) as

$$y_t = \mu + \sum_{i=0}^{\infty} \Psi^g(i) u_{t-i}^g. \tag{5}$$

Henceforth, we can interpret $u_t^g$ as the innovation vector in the generalized impulse response analysis, where a single element receives a unity shock and all others remain at zero. Implicitly, we are assuming that this is an orthogonal and exogenous shock vector. However, due to the distribution of $u_t^g$ with the inverse correlation matrix $\Omega$, this step results in an approximation error. The approximation error can be linked to the partial correlation matrix of $u_t$, and it is generally bigger when more entries are partially correlated.[3]

Note that the $h$-period forecast error variance of variable $i$ explained by innovations in variable $j$ is $\left(\psi_{ij}^g(h)\right)^2$. Then, the effect of a generalized impulse of variable $j$ at time $t$ on the $H$-step-ahead forecast error variance of variable $i$ is

$$MSE\left[y_{i,t+H-1}\middle|\{u_{t+h}^g = e_j\}_{h=0}^{H-1}\right] = \sum_{h=0}^{H-1}\left(\psi_{ij}^g(h)\right)^2. \tag{6}$$

The $H$-step-ahead forecast error variance contributions from all variables to $i$ is just its mean squared error (MSE):

$$MSE[y_{i,t+H-1}] = \left(\sum_{h=0}^{H-1}\left(\Phi_h \Sigma \Phi_h'\right)\right)_{ii}. \tag{7}$$

Pesaran and Shin (1998) divide (6) by (7) and get a table showing the contributions from innovations in variable $j$ to the $H$-step-ahead forecast error variance of variable $i$. Like Diebold and Yılmaz (2014), we denote the $H$-step-ahead generalized FEVD as $D^{gH} = [d_{ij}^{gH}]$ with entries

$$d_{ij}^{gH} = \frac{MSE\left[y_{i,t+H-1}\middle|\{u_{t+h}^g = e_j\}_{h=0}^{H-1}\right]}{MSE\left[y_{i,t+H-1}\right]}. \tag{8}$$

Note that the numerator implicitly shocks single entries of $u_t^g$ and the denominator shocks single entries of $u_t$. In other words, the "generalized" FEVD approximates shocks with $u_t^g$ and is only accurate if $\Omega = I_N$; i.e., if $\Sigma$ is diagonal. Optionally, Diebold and Yılmaz (2014) row-normalize these tables for a cleaner network interpretation. However, row normalization distorts the entries such that the distribution of the estimation errors gets more complex. Thus, if not explicitly stated, we do not perform this step.

[3]For a discussion of $\Omega$ see Raveh (1985).

## 2.2 Estimating Large Forecast Error Variance Decompositions

### 2.2.1 Regularizing Vector Autoregressive Coefficients

As in Kascha and Trenkler (2015), we first transform the VAR setup such that our coefficient matrix $A$ can be estimated in vector form

$$y = (Z' \otimes I_N)\,\beta + u, \tag{9}$$

where $y = vec([y_1, ..., y_T])$, ${Z^0_{t-1}}' = (y'_{t-1}, ..., y'_{t-p})'$, $Z_{t-1} = (1, {Z^0_{t-1}}')'$, $Z = [Z_0, ..., Z_{T-1}]$, $\beta = vec(A)$ and $u = vec([u_1, ..., u_T])$. We set $X := (Z' \otimes I_N)$ to obtain the general regression form. For a general regression, the ordinary least squares (unpenalized) estimator reads

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||^2.$$

Here $||\cdot||^2$ denotes the square of the Frobenius norm. Based on this objective function, we aim to regularize the coefficient matrix $A$. On this account, we consider elastic-net regularization that comprises the extreme cases of LASSO and ridge regression. Additionally, we introduce a new regularization target that enforces sparsity on the long-run dependencies.

**(Adaptive) elastic-net, LASSO and ridge regression** We outline the most general concept following Zou and Zhang's (2009) adaptive elastic-net. This penalized estimator is a compound of the general concepts of the elastic-net and adaptive LASSO. In particular, it simultaneously shrinks and selects entries in the coefficient matrices and, moreover, has the oracle property, which ensures optimal large sample performance. A comprising definition of the adaptive elastic-net estimator class is

$$\hat{\beta}_{AEnet} = \underset{\beta}{\operatorname{argmin}} \left[ ||y - X\beta||^2 + \lambda \sum_{i=1}^{N^2 p} w_i \left( \alpha|\beta_i| + (1-\alpha)\frac{1}{2}\beta_i^2 \right) \right], \tag{10}$$

where $w_i = |\hat{\beta}_{i,ini}|^{-\gamma}$ is an initial guess with $\gamma > 0$ and $\lambda$ is a tuning parameter that controls the strength of the elastic-net penalty, and must be chosen by the researcher, e.g., via cross-validation (CV). Note that the original paper proposed to only use the weights on the LASSO penalty. However, similar to Demirer et al. (2017), we put the weight before the shrinkage penalty and use the `glmnet` routine from Friedman et al. (2010).[4]

[4]The `glmnet` routine is available for many programming languages. We employ the implementation in MATLAB by Qian et al. (2013).

The regression in (10) is an enhanced version of the penalty regression and, thus, generalizes many other concepts. For example, the elastic-net penalty with $\alpha \in (0,1)$ combines the LASSO and the ridge estimator and inherits the desirable properties of both; e.g., it removes the degeneracy of the LASSO estimator caused by extreme correlations while still performing variable selection. Moreover, in the case of a large $N$ and a small $T$, LASSO selects at most $NT$ nonzero entries before it saturates. The mixture with the ridge penalty eradicates this behavior. In a nutshell, the absolute penalty term automatically selects variables while the quadratic penalty shrinks and stabilizes the solution paths (see Zou and Zhang, 2009).

Now, choosing $w_i = 1$ gives the naive elastic-net,[5] $\alpha = 1$ the adaptive LASSO, $\alpha = 0$ the ridge regression, and $\alpha = w_i = 1$ the classical LASSO from Tibshirani (1996). In Section 3, we compare the performances of the LASSO, the ridge regression and the adaptive elastic-net in a simulation of FEVDs. If not stated differently, we use $w_i = |\hat{\beta}_{i,OLS}|^{-1}$.

**Regularizing long-run effects** While most regularization methods assume sparsity in either the regression coefficients or the covariance matrix, economic setups sometimes motivate independence between variables at all time horizons. As FEVDs represent the overall dependency with different response times between variables, the idea of sparsity may also apply here. In this subsection, we briefly introduce a method that models sparsity in the long-run dependency.

Recall that we can use a VAR(1) without loss of generality since it represents any VAR(p) in companion form. Then, the moving average matrices are $\Phi_i = A^i$. If economic theory implies that only a few variables affect each other, we ideally also presume sparsity in the lagged responses. For example, if we assume sparsity in the long-run responses, then we should regularize all $\Phi_k$'s. This regularization proves to be difficult since the power of a matrix is a complex function of the coefficient matrix $A$.

To overcome this issue, we take the forecast error, i.e., the response to a one-standard-deviation impulse. The long-run (lagged) response of an impulse is

$$\begin{aligned} FE(H) &= \sum_{h=0}^{H-1} \Phi_h = \sum_{h=0}^{H-1} A^h, \\ \lim_{H\to\infty} FE(H) &= \sum_{h=0}^{\infty} A^h = (I_N - A)^{-1}, \end{aligned} \tag{11}$$

where the last equation holds due to the stability condition and is the result of the geometric series. If there is no spillover of one variable to another, we assume that the

[5] Zou and Zhang (2009) rescale the naive estimator by $(1 + (1-\alpha)\lambda/T)$. Similar to Friedman et al. (2010), we drop this distinction.

respective entry in $FE(\infty) = (I_N - A)^{-1}$ is zero.

It is evident that zeros in the forecast error most likely imply that the respective forecast error variance — i.e., the element-wise squared version — is also zero. Thus, to impose sparsity on the spillover network, this matrix is a potential regularization target. Take (11) and plug it into the model (1):

$$\begin{aligned} y_t &= \nu + (I_N - FE(\infty)^{-1})y_{t-1} + u_t, \\ \Delta y_t = y_t - y_{t-1} &= \nu - FE(\infty)^{-1}y_{t-1} + u_t, \\ y_{t-1} &= FE(\infty)\nu + FE(\infty)(-\Delta y_t) + FE(\infty)u_t \end{aligned} \tag{12}$$

Estimating (12) as a penalized regression permits us to regularize $FE(\infty)$ and with $A = I_N - FE(\infty)^{-1}$ we can back out the autoregressive coefficients. Whereas this version quite likely performs worse if there is sparsity only in $A$, it may be better if the FEVD is sparse itself. Henceforth, we denote the LASSO regularization on this target as the geometric regularization.

### 2.2.2 Regularizing the Covariance Matrix

Recall that FEVDs are functions of the coefficient and covariance estimates. In particular, all forecast errors include the impact-period response; i.e., some decomposition of the covariance matrix. Even though the generalized variance decomposition is not a decomposition of the covariance matrix estimate itself, it uses this estimate as an input. Since the covariance matrix estimate also suffers under high dimensionality, we regularize the covariance matrix to obtain better inputs for the FEVDs. A variety of methods originated from the literature on the regularization of covariance matrices; we introduce three of them. We assume that the innovation series $u_t$ is known. That is, we assume that the regularized estimation of the VAR coefficients works reasonably well. Note that the following overview is formulated in terms of the mean zero random vector $\mathbf{u}$. That is, before applying any of the following methods, it is advisable to demean the variables. In our simulation and application, observations of $\mathbf{u}$ will be the innovation series $u_t$ corresponding to the regression residuals $\hat{u}_t$ in the first stage.

**Adaptive thresholding** Thresholding methods are designed for sparse covariance matrices and they mimic the ideas of shrinkage and selection of entries in the sample covariance matrix. Assume an $N$-variate random vector $\mathbf{u} = (\mathbf{u}_1, ..., \mathbf{u}_N)'$ with covariance matrix $\Sigma = [\sigma_{ij}]_{N \times N}$. Furthermore, assume an *i.i.d.* random sample $\{u_1, ..., u_T\}$ from the distribution of $\mathbf{u}$. The aim is to estimate the covariance matrix $\Sigma$ with big $N$ and

small $T$. Start with the sample covariance matrix

$$\hat{\Sigma} = [\hat{\sigma}_{ij}]_{N\times N} := \frac{1}{T-1}\sum_{t=1}^{T} u_t u_t'.$$

Further, define the variance of the sample covariance's entries as

$$\theta_{ij} := \mathrm{Var}\left(\mathbf{u}_i\mathbf{u}_j\right) = E\left(\left(\mathbf{u}_i\mathbf{u}_j - \sigma_{ij}\right)^2\right). \tag{13}$$

We can now interpret the sparse covariance estimation as a mean vector estimation. That is, an individual entry can be described as

$$\frac{1}{T}\sum_{t=1}^{T} u_{i,t}u_{j,t} = \sigma_{ij} + \sqrt{\frac{\theta_{ij}}{T}}z_{ij}, \tag{14}$$

with $z_{ij}$ asymptotically standard normal. On this basis, it is straightforward to create an individual threshold for each entry of the covariance matrix. Yet, the variability of an individual entry, $\theta_{ij}$, needs to be estimated with its sample counterpart $\hat{\theta}_{ij} = T^{-1}\sum_{t=1}^{T}\left[u_{i,t}u_{j,t} - \hat{\sigma}_{ij}\right]^2$. Note that we do not subtract the sample mean since $u_t$ is demeaned. Cai and Liu (2011) propose calculating entry-wise thresholds by

$$\lambda_{ij} := \lambda_{ij}(\delta) = \delta\sqrt{\frac{\hat{\theta}_{ij}\log N}{T}}, \tag{15}$$

where the regularization parameter $\delta > 0$ can be selected with CV or, as the authors suggest, be set to 2. The function $s_\lambda(z)$, which we discuss in more detail below, describes the thresholding rule applied to the entries of the sample covariance. Finally, for a given $s_\lambda(z)$, the adaptive (entry dependent) threshold estimator

$$\hat{\Sigma}^{AT} = [\hat{\sigma}_{ij}^{AT}]_{N\times N} = [s_{\lambda_{ij}}(\hat{\sigma}_{ij})]_{N\times N} \tag{16}$$

allows for different threshold levels for the entries and incorporates the variability of the entries.

Note that the researcher can choose the thresholding functions $s_\lambda(\cdot)$ to determine the type of the penalty. In particular, we consider the adaptive-thresholding rule $s_\lambda(\hat{\sigma}_{ij}) = \hat{\sigma}_{ij}(1-|\lambda/\hat{\sigma}_{ij}|^\eta)_+$, with $(z)_+ = max\{z, 0\}$.[6] Unlike before, $\lambda$ is a function described by (15) and, thus, is not freely selectable in the adaptive-threshold estimator. Here, $\delta$ determines the degree of the penalization and has to be selected by the econometrician. For a specific

[6]This rule incorporates soft thresholding for $\eta = 1$: $s_\lambda(\hat{\sigma}_{ij}) = \mathrm{sgn}(\hat{\sigma}_{ij})(|\hat{\sigma}_{ij}| - \lambda)_+$. Cai and Liu (2011) describe this rule for $\eta \geq 1$. We will relax this condition and include values between zero and one in later steps.

class of thresholding functions, the estimator achieves optimal convergence and performs better than the universal thresholding estimator, which uses the same threshold for all entries. This class must satisfy some conditions as described in Appendix A.

Cai and Liu (2011) suggest a thresholding rule that applies to all entries in the covariance matrix. In simulations, we experienced zero entries on the diagonal of the threshold estimates. Since this behavior contradicts the idea of covariances, we treat diagonal values differently. We describe the details in Appendix B.

**Ledoit and Wolf's (2004) bona-fide shrinkage estimator** The estimator of Ledoit and Wolf (2004) is well-conditioned (inverting it does not amplify estimation errors) and more accurate than the sample covariance matrix. In particular, the estimator is the optimal convex linear combination of the sample covariance and the identity. Optimality is achieved asymptotically with respect to a squared loss function. We consider this regularization method, since it is easy to compute and *bona fide*; i.e., it is not affected by the choice of a regularization term and does not require additional knowledge. The Ledoit-Wolf shrinkage estimator reads as follows:

$$\hat{\Sigma}^{LW} = \frac{b^2}{d^2} m I_N + \left(1 - \frac{b^2}{d^2}\right) \hat{\Sigma}, \tag{17}$$

with

$$\begin{aligned} m &= \operatorname{tr}(\hat{\Sigma}) N^{-1}, \qquad d^2 = ||\hat{\Sigma} - m I_N||^2, \\ b^2 &= \min \left[ T^{-2} \sum_{t=1}^{T} ||u_t u_t' - \hat{\Sigma}||^2, d^2 \right], \end{aligned}$$

where $\hat{\Sigma}$ denotes the sample covariance matrix and $m$ is the average of the diagonal values of the sample covariance matrix. In contrast to Ledoit and Wolf (2004), we use the scaled sample covariance matrix; i.e., we divide the sum by $T - 1$ instead of $T$. This scaling is negligible for sufficiently large $T$.

The estimator is a linear shrinkage estimator, which optimally mixes the "all-bias no-variance" estimator $mI_N$ with the "all-variance no-bias" estimator $\hat{\Sigma}$. The term $b^2/d^2$ automatically assigns more weight to $mI_N$ if the variance of the sample's second order, measured by $b^2$, is large. Thus, similar to the shrinkage in James and Stein (1992), $\hat{\Sigma}^{LW}$ trades off the bias and variance of the estimator to minimize the MSE. Note that the shrinkage weight $\delta = b^2/d^2$ can also be chosen with CV.

**GLASSO** Friedman et al. (2008) propose estimating sparse graphs by penalizing the inverse covariance matrix. That is, they estimate the inverse covariance with a LASSO penalty. Let $\Theta = \Sigma^{-1}$ and optimize

$$\hat{\Theta}^{GLASSO} = \underset{\Theta \geq 0}{\operatorname{argmax}} \left\{ \log \det \Theta - \operatorname{tr}(\hat{\Sigma}\Theta) - \delta \, ||\Theta||_1 \right\}, \tag{18}$$

over all nonnegative definite matrices $\Theta$, where $|| \cdot ||_1$ denotes the $L_1$ norm, i.e., the sum of the absolute values of the entries. As before, $\hat{\Sigma}$ is the sample covariance. The first two terms describe the multivariate Gaussian likelihood, and the latter is the LASSO penalty, which selects entries in $\Theta$ and sets others to zero. The matrix $\hat{\Theta}$, which minimizes this objective function, is the respective estimator of the inverse covariance matrix. Since the approach intends to estimate undirected graphical models, they call this estimation the graphical LASSO or, in short, GLASSO.

The fact that the penalization is on the inverse covariance makes this approach particularly appealing in the setup of FEVDs. For example, Barigozzi and Brownlees (2013) highlight the relation of the inverse covariance (concentration matrix) to partial correlations. In particular, if entry $[\Sigma^{-1}]_{ij}$ in the inverse covariance is zero, then variables $i$ and $j$ are conditionally uncorrelated. In economic setups, it is often plausible to assume sparsity in the partial correlations rather than in the overall correlation structure. In other words, it is easier to rule out direct than indirect effects.

## 2.3 Data-Driven Choice of Regularization Parameters $\lambda$ and $\delta$

Except for the Ledoit-Wolf shrinkage estimator, all of the above estimators require the choice of regularization parameters.[7] In particular, techniques related to the elastic-net and the graphical LASSO require the choice of $\lambda$ and the adaptive-thresholding rule requires the choice of $\delta$. Technically, one can also choose the shrinkage weight in Ledoit-Wolf, but since the authors complement their estimator with an optimized weight, this step is optional. To validate the goodness of fit of these parameters, we resort to statistical learning methods. Specifically, we use CV in our application.

In general terms, model validation procedures such as CV divide the sample into training and test data.[8] Given fixed hyperparameters, these methods estimate the remaining parameters from the training data and validate their performance on the test data. That is, the quality of the estimate from the training dataset is assessed by its ability to explain the test dataset. The model selected after validation is the version that best explains the test sample. A common practice for small datasets is $K$-fold CV, which partitions the

[7]The literature often refers to these parameters as hyperparameters as they must be fixed before estimating the other parameters.

[8]The statistical learning literature uses the term "validation set" to denote the hold-out sample in each CV fold. In econometric applications, however, it is common to denote the subset of the data used for hyperparameter tuning as the "test data." We stick to the latter terminology throughout our paper.

sample into $K$ equally sized test samples.[9] For each of these $K$ samples, CV trains the estimator on the remaining training data and then validates it on the test sample. A loss function such as the MSE or the negative log-likelihood quantifies performance.

While CV works well for most setups, it is worth highlighting the predicament that occurs in high dimensions. Since its main idea is dividing the sample into test and training data, this split decreases the number of observations that are usable for estimation. The estimator faces a different degree of estimation uncertainty and, thus, CV tends to choose a stronger penalization term. Consequently, leave-one-out CV is the preferred choice since it uses $T-1$ observations as training data, which is close to the full sample size. On the other hand, a lower $K$ (e.g., 10) is expected to estimate the prediction error better since it averages over more-diverse training sets (Hastie et al., 2009, Chapter 7.12). Hence, there is a trade-off between a good training sample and a reasonable estimate of the prediction error.

While CV is valid in an *i.i.d.* setup, it lacks theoretical justification for dependent data, such as time series regressions. Take for example a two-fold CV, which divides the sample into two sets. In this case, the observations at the boundaries show up in both samples due to autoregressive effects. Therefore, an outlier may affect both the dependent and independent variables. Leaving out observations between the samples circumvents that issue but leads to a significant loss of observations with random sampling. Thus, it is common practice to use block sampling, as described by Figure C.11 in Appendix C. Bergmeir et al. (2018) show that the $K$-fold version is permissible for a purely autoregressive model with uncorrelated innovations. Consequently, we recommend sticking to $K$-fold CV with nonrandom sampling.

Unlike CV for coefficient estimates, there exists no dominant established version for covariance estimation. To understand why this is the case, it is worth taking a look at the loss functions. CV for the coefficient estimates, here $\beta$, minimizes the MSE of the implied residuals. This loss function is straight forward since it is also the minimization target in the estimation step. However, the validation of the covariance matrix proves to be more abstract due to the unobservability of the covariance matrix. Cai and Liu (2011) propose using the sample covariance estimate of the test sample to validate the goodness of fit. The loss function for the $k$th test sample is

$$\ell_k(\delta) = ||\widehat{\Sigma}_{\{1:T\}\setminus k}(\delta) - \widehat{\Sigma}_k||^2,$$

where the first term of the difference is the training data estimate of $\Sigma$ with penalization parameter $\delta$. The second term is the sample covariance of the test data sample. Following the motivation of this paper, however, this loss function is imprecise since the sample

[9] See Appendix C for an illustration of the procedure. The special case of leave-one-out CV is obtained if $K$ equals the sample size, validating the estimator on a single observation.)

covariance is a bad estimate of the test sample. Thus, we prefer to directly validate the trained estimate $\widehat{\Sigma}_{\{1:T\}\setminus k}(\delta)$ on the squared observations $u_t u_t'$. Naturally, we want to use a similar loss function as in the optimization of the estimator. For example, we can use a likelihood-based loss function similar to the GLASSO approach. Unfortunately, we experienced numerical issues with this loss function since the likelihood is not always well defined for the regularized version.

Alternatively, we propose to use a modified version of Cai and Liu (2011). For the $k$th test sample, it is calculated as

$$\ell_k(\delta) = \frac{1}{T_k} \sum_{t \in k} ||\widehat{\Sigma}_{\{1:T\}\setminus k}(\delta) - u_t u_t'||^2,$$

where $T_k$ denotes the sample size of $k$. In contrast, this loss function shows the mean of the distance between the squared observations and the trained estimate.

For computational reasons, econometricians are limited to a relatively small set of hyperparameters. Namely, there are $k$ estimations for each set of hyperparameters. For example, a $k$-fold CV for 100 different values of both $\lambda$ and $\alpha$ requires $100^2 k$ estimations. While large CVs are achievable with today's computation power, it is advisable to limit the grid search to a neighborhood of the presumed minimum loss. Thus, we aim to produce grid values centering around the minimum MSE. Similar to Kascha and Trenkler (2015), we increase the candidate hyperparameters linearly on a log scale between data-driven minimum and maximum values.[10]

At this point, it is worth mentioning an obvious extension. Since CV selects penalty terms on predefined sample splits, we can also plug in different regularization techniques for the estimators. CV then selects the estimator with the best predictive power. We apply this procedure to our empirical setting in Section 4.

# 3 Simulation Study

## 3.1 Data Generating Processes

In the context of FEVDs, we believe that sparsity emerges with increasing dimension $N$. However, it remains unclear whether sparsity appears in the VAR coefficients, the innovation covariance matrix, the FEVD, or in all of them. Thus, we introduce various data generating processes (DGPs) and hope to address the most relevant issues. Although this simulation study is limited to VAR(1) models, it also extends to higher lag orders in companion form. Estimating models with additional lags will, however, aggravate

[10] Friedman et al. (2010) provide formulas to set the minimum and maximum hyperparameters for general elastic-nets.

the high-dimensionality problem. To resemble real-world observational data, the DGPs include sparse coefficients in the VAR matrix $A$ and innovation covariance $\Sigma$.

*DGP 1:* White noise process (in the observations)

$A = \mathbf{0}_{N \times N}$
$\Sigma = I_N$

*DGP 2:* Diagonal VAR coefficient (auto-correlation with no spillovers)

$A = 0.5 I_N$
$\Sigma = I_N$

*DGP 3:* Diminishing-diagonal VAR coefficient (only approximately sparse)

$A$ is a banded diagonal matrix with entries $a_{ij} = 0.3^{|i-j|+1}$
$\Sigma = I_N$

*DGP 4:* Random graph (without ordering)

$A$ is a random sparse matrix. That is, it has entries with probability $P(a_{ij} \neq 0) = \tau$. $\tau = 1/\sqrt{N}$ denotes the degree of density. Nonzero entries are temporarily set to one. Then, $A$ is rescaled such that its maximum eigenvalue is 0.5 in modulus.
$\Sigma = SD\,S'$. Here, $S$ is the long-run dependency spanned by $A$, i.e., $S = (I_N - A)^{-1}$. $D = diag(1, ..., 2)$ is an $N \times N$ diagonal matrix with values spaced equally between 1 and 2.

*DGP 5:* Block-diagonal FEVD (sparse FEVD)

This DGP mimics a network relation for time series with a sparse FEVD. First, a diagonal block structure with $1/\tau$ equal-sized quadratic blocks is chosen for $A$. $\tau = 1/\sqrt{N}$ denotes the degree of density. Nonzero entries are temporarily set to one on the diagonal and 0.5 on the off-diagonal. Then, $A$ is rescaled such that its maximum eigenvalue is 0.5 in modulus.
$\Sigma = SD\,S'$. Here, $S$ is the long-run dependency spanned by $A$, i.e., $S = (I_N - A)^{-1}$. $D = diag(1, ..., 2)$ is an $N \times N$ diagonal matrix with values increasing from 1 to 2.

Note that *DGPs 1*, *2*, and *5* always produce a sparse FEVD and *DGPs 3* and *4* do not. Moreover, all but *DGP 4* are deterministic and *DGPs 1-3* have the identity as the covariance matrix. *DGP 4* has 25 different random realizations in the simulation. The methods' performances are compared for FEVDs with forecast horizon 10.

## 3.2 Comparison of Different Regularization Methods

This section analyzes the relative performance gain of using regularization methods to estimate FEVDs. That is, we compare the estimation errors of the various regularized versions to the OLS plus the sample covariance case. In particular, we run simulations for the aforementioned DGPs and calculate the Frobenius norm of the regularized versions versus the case using OLS plus the sample covariance: $||\widehat{D}^{gH}_{reg} - D^{gH}||/||\widehat{D}^{gH}_{OLS} - D^{gH}||$

with $H = 10$. We compare the performance gains over $N = \{50, 150, 250\}$ and $T = \{75, 175, 500\}$. Since the estimation is a two-step procedure, we split the simulation into two parts. Note that OLS breaks down for $N > T$ such that we are not able to calculate any value for these cases.[11] First, we regularize $A$ paired with the sample covariance for $\Sigma$. We compare the ridge, LASSO, adaptive elastic-net, and geometric long-run regularization. The latter uses the LASSO penalty only, such that it performs variable selection. The penalty parameters $\lambda$ are chosen such that they minimize the respective norm $\lambda^* = \text{argmin}_\lambda ||\widehat{D}^{gH}_{reg}(\lambda) - D^{gH}||$. Thus, the values show the best possible performance gain.

| ***DGP1*** | | Ridge | | | LASSO | | | AENET | | | Geometric | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N\backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 21.5% | **22.5%** | **22.6%** | 21.6% | 22.7% | 22.6% | **21.5%** | 22.6% | 22.6% | 22.4% | 23.1% | 22.8% |
| 150 | | 12.3% | **13.8%** | | **12.3%** | 13.8% | | 12.3% | 13.8% | | 12.6% | 13.9% |
| 250 | | | 10.8% | | | 10.8% | | | **10.7%** | | | 10.9% |
| ***DGP2*** | | Ridge | | | LASSO | | | AENET | | | Geometric | |
| $N\backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 21.5% | **22.5%** | **22.6%** | 21.6% | 22.7% | 22.6% | **21.5%** | 22.6% | 22.6% | 22.4% | 23.1% | 22.8% |
| 150 | | 12.3% | **13.8%** | | **12.3%** | 13.8% | | 12.3% | 13.8% | | 12.6% | 13.9% |
| 250 | | | 10.8% | | | 10.8% | | | **10.7%** | | | 10.9% |
| ***DGP3*** | | Ridge | | | LASSO | | | AENET | | | Geometric | |
| $N\backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 23.6% | 23% | **23.9%** | **23.3%** | **22.9%** | 25.0% | 23.6% | 23.0% | 24.5% | 29.8% | 29.9% | 35.4% |
| 150 | | 13.7% | 14.5% | | **13%** | **14.1%** | | 13.6% | 14.2% | | 17.2% | 19.2% |
| 250 | | | 11.7% | | | **11.1%** | | | 11.2% | | | 15.2% |
| ***DGP4*** | | Ridge | | | LASSO | | | AENET | | | Geometric | |
| $N\backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 27.9% | **34.6%** | **47.8%** | 27.6% | 35.1% | 48.7% | **27.3%** | 34.8% | 48.3% | 35.7% | 51.0% | 87.3% |
| 150 | | 13.9% | 17.8% | | 13.7% | **17.8%** | | **13.7%** | 17.8% | | 16.0% | 23.5% |
| 250 | | | 12.3% | | | 12.3% | | | **12.2%** | | | 14.4% |
| ***DGP5*** | | Ridge | | | LASSO | | | AENET | | | Geometric | |
| $N\backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 30.7% | **39.2%** | **55.7%** | 30.5% | 39.6% | 57.8% | **30.3%** | 39.3% | 57.5% | 43.3% | 62.7% | 111.5% |
| 150 | | 16.7% | **24.7%** | | 16.4% | 24.9% | | **16.4%** | 24.8% | | 24.5% | 45.3% |
| 250 | | | 12.4% | | | **12.3%** | | | 12.3% | | | 15.1% |

Table 1: Simulation results for the regularization of $A$ paired with the sample covariance. $||\widehat{D}^{gH}_{reg} - D^{gH}||/||\widehat{D}^{gH}_{OLS} - D^{gH}||$ are shown for different $N$s and $T$s with 500 Monte Carlo repetitions. *DGPs 4* and *5* have 25 different random realizations of $A$ and $\Sigma$. Ridge denotes the ridge regression and AENET denotes the adaptive elastic-net.

Table 1 contains the simulation results for the regularization of $A$ for 500 Monte Carlo repetitions. First, it is evident that regularization results in a large overall efficiency gain. *DPGs 1-3* all show similar gains for all estimators. The best performance gain for all DGPs and regularization methods is at $N = 250$. However, for $N = 50$ and $T = 500$ we still observe a remarkable efficiency gain. That is, the best regularized estimators achieve reductions in the norms to the true values to $10.7 - 23.9\%$ for *DGPs 1-3* and $12.2 - 55.7\%$ for *DGPs 4-5*. Only the geometric regularization method performs worse than OLS plus the sample covariance for $N = 50$ and $T = 500$.

Surprisingly, all traditional regularization methods perform similarly well. There is no

[11] In general, regularization methods are not limited by $N > T$.

clear winner among the estimators since the differences in performance are marginal. If at all, the performance of the ridge estimator appears better for cases in which $T$ is relatively large. The variable selection methods LASSO and adaptive elastic-net perform better for simulations where $T$ is close to $N$. Further, the comparison of the ridge estimator, the LASSO and the adaptive elastic-net for specific *DGPs* reveals that neither is dominant under particular circumstances. Finally, the geometric regularization performs slightly worse for all *DGPs*.

As a second step, we compare regularization methods for $\Sigma$. That is, we calculate the residuals using the adaptive elastic-net estimate and construct the FEVD with the (regularized) estimate of $\Sigma$. The adaptive elastic-net penalty parameter is chosen to minimize the Frobenius norm of the difference between the estimate matrix $\hat{A}$ and the true parameter matrix $A$: $\lambda^* = \text{argmin}_\lambda ||\hat{A}(\lambda) - A||$. We use the soft-thresholding rule for the adaptive-threshold estimator.

| ***DGP1*** | Sample-Cov | | | Threshold | | | Ledoit-Wolf | | | GLASSO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N \backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 21.6% | 22.5% | 22.6% | **0.0%** | **0.0%** | **0.0%** | 0.1% | **0.0%** | **0.0%** | **0.0%** | **0.0%** | **0.0%** |
| 150 | | 12.3% | 13.8% | | **0.0%** | **0.0%** | | **0.0%** | **0.0%** | | **0.0%** | **0.0%** |
| 250 | | | 10.8% | | | **0.0%** | | | **0.0%** | | | **0.0%** |
| ***DGP2*** | Sample-Cov | | | Threshold | | | Ledoit-Wolf | | | GLASSO | | |
| $N \backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 21.6% | 22.5% | 22.6% | **0.0%** | **0.0%** | **0.0%** | 0.1% | **0.0%** | **0.0%** | **0.0%** | **0.0%** | **0.0%** |
| 150 | | 12.3% | 13.8% | | **0.0%** | **0.0%** | | **0.0%** | **0.0%** | | **0.0%** | **0.0%** |
| 250 | | | 10.8% | | | **0.0%** | | | **0.0%** | | | **0.0%** |
| ***DGP3*** | Sample-Cov | | | Threshold | | | Ledoit-Wolf | | | GLASSO | | |
| $N \backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 24.3% | 23.1% | 27.4% | **3.4%** | **8.5%** | **19.8%** | 3.6% | 8.5% | 19.9% | 3.4% | 8.5% | 19.8% |
| 150 | | 14.4% | 14.8% | | **2.8%** | **7.7%** | | 2.8% | 7.7% | | 2.8% | 7.7% |
| 250 | | | 11.5% | | | **4.6%** | | | 4.6% | | | 4.6% |
| ***DGP4*** | Sample-Cov | | | Threshold | | | Ledoit-Wolf | | | GLASSO | | |
| $N \backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 29.2% | 38.6% | 58.1% | 12.5% | 27.5% | **58%** | **12.4%** | **26.5%** | 59.5% | 12.9% | 27.5% | 58.4% |
| 150 | | 14.2% | 20% | | **3.9%** | 11.8% | | 3.9% | **11.8%** | | 4% | 12.1% |
| 250 | | | 13% | | | 5.1% | | | **5.1%** | | | 5.2% |
| ***DGP5*** | Sample-Cov | | | Threshold | | | Ledoit-Wolf | | | GLASSO | | |
| $N \backslash T$ | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 | 75 | 175 | 500 |
| 50 | 32.3% | 41.5% | 61.9% | 23.3% | 38.4% | **61.5%** | 25.5% | 46.9% | 84.6% | **23.2%** | **38.1%** | 62.4% |
| 150 | | 18.1% | 28.8% | | 12.3% | 26.4% | | 14.5% | 34.9% | | **12%** | **26.3%** |
| 250 | | | 12.8% | | | 5.7% | | | 6.2% | | | **5.6%** |

Table 2: Simulation results for the regularization of $\Sigma$, paired with the best-performing adaptive elastic-net estimator. Results are divided by the estimation uncertainty of the OLS estimator with the sample covariance matrix. $||\widehat{D}^{gH}_{reg} - D^{gH}||/||\widehat{D}^{gH}_{OLS} - D^{gH}||$ are shown for different $N$s and $T$s with 500 Monte Carlo repetitions. *DGP 4* has 25 different random realizations of $A$ and $\Sigma$.

Table 2 shows simulations with different regularizations. The first estimator is the sample covariance matrix and sets the benchmark for the performance of the covariance regularization. Again, we measure the performance of the regularization methods with the norm of the estimated FEVD to the true values. The most salient result is the 0% norm for *DGPs 1-2*. In particular, the FEVDs appear perfectly estimated, since the data generating covariance matrix is the identity. Hence, the strongest penalization can find this matrix. For example, shrinkage towards a diagonal matrix as in the Ledoit-

Wolf estimator approximately estimates the identity. Similarly, the adaptive-threshold and GLASSO estimators both select diagonal entries in the covariance matrix. Note, however, that the perfect estimation of the coefficient matrix in the first step is also a precondition for the efficient regularization of the innovation covariance matrix and, thus, for the perfect fit of the overall FEVD.

In *DGP 3*, the elastic-net is not able to perfectly estimate this matrix since it is not sparse. Thus, the estimation of the overall FEVD is not completely perfect. However, it appears that sometimes the methods perfectly estimate the covariance matrix, which is again the identity. The performance is similar for all the regularization methods. Different performances for the methods are observable for the non-diagonal covariance matrices in *DGPs 4-5*. A significantly increased performance is observable for all combinations of $N$ and $T$ except for $N = 50$ and $T = 500$. For this case, the sample covariance performs reasonably well such that its regularization does not enhance estimation accuracy. Again, we do not find pronounced differences in the performances between these estimators.

In summary, we find that the regularization of the two estimation steps leads to a significant performance gain in the estimation of FEVDs. The best-performing estimator in our simulations is the ridge regression combined with the Ledoit-Wolf estimator, although by a thin margin. Both estimators slightly outperform the others and are also easy to apply. However, since we already have to find the penalization hyperparameter, it is advisable to use CV techniques to validate these methods' performance for specific data.

## 3.3 Bias, Variance, and Edge Detection

First, we take a look at the mean of the entries and the norms for regularized and unregularized FEVDs. For this, we construct time series with *DGP 5*, $N = 100$ and $T \in [100, 500]$. The analysis compares OLS and the adaptive elastic-net, combined with the sample covariance matrix and the GLASSO estimate, respectively. We select the hyperparameters based on the best performance (minimization of the Frobenius norm to the true FEVD). Note that this procedure requires knowledge about the true parameters, such that it is applicable in a simulation exercise but cannot be performed in an empirical application. The left-hand panel in Figure 2 shows the average mean distance, $N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} mean(\widehat{D}_{ij}^{gH} - D_{ij}^{gH})$, which allows us to interpret the results as the biases of the FEVD estimates. The right-hand panel of Figure 2 then shows the average variance $N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} var(\widehat{D}_{ij}^{gH})$. Note that this analysis relates to the "bias-variance" trade-off for estimators as regularization methods generally sacrifice unbiasedness to achieve a lower variance. If balanced optimally, this step leads to a reduction in prediction errors overall.

The left-hand panel depicts the magnitude of the bias for small $T$ (approaching $N =$

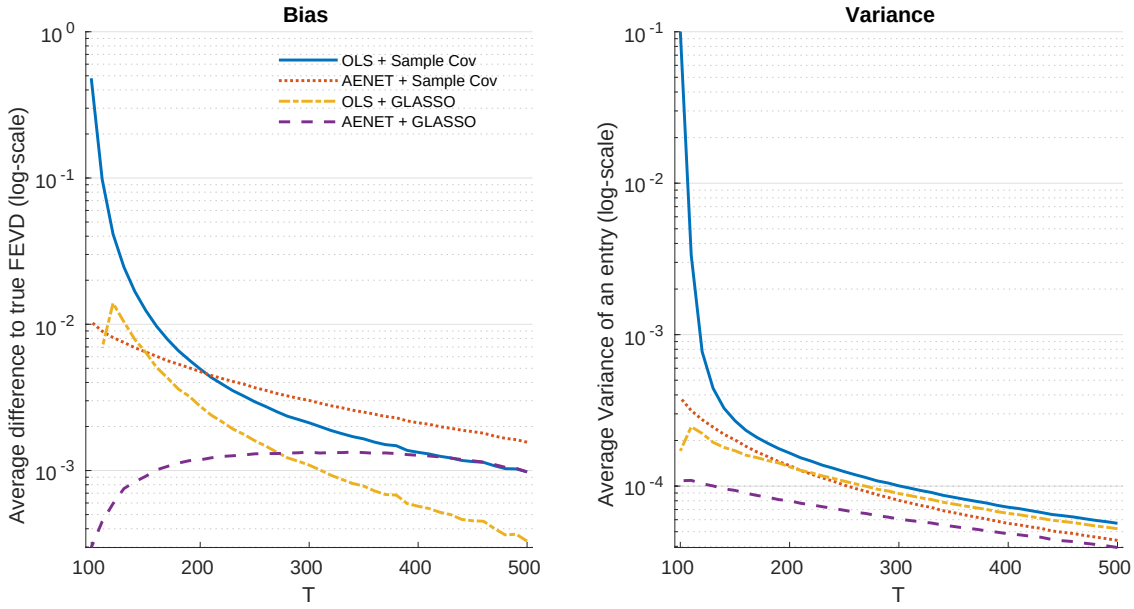


Figure 2: Simulation results for 500 Monte Carlo repetitions of *DGP 5* and $N = 100$. The left-hand panel shows the average mean difference of the estimates to the true values $N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} mean(\widehat{D}_{ij}^{gH} - D_{ij}^{gH})$. The figure shows the bias of the resulting estimate for the respective curves. The right-hand panel shows the respective average variance $N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} var(\widehat{D}_{ij}^{gH})$. The sample size $T$ is on the x-axis. AENET denotes the adaptive elastic-net.

100 from the right). We see that on average all estimators are positively biased, with the non-regularized estimator (blue solid line) profoundly overestimating entries in the FEVDs. That is, this estimator faces a strong positive bias the closer $T$ is to $N$. Perhaps surprisingly, the adaptive elastic-net (red dotted curve) already diminishes the bias for small $T$ by a margin (almost by a factor of 100 for $T = 100$). While we would have expected that regularization methods mostly trade off variance for bias, it appears that for $T < 200$ the adaptive elastic-net also performs better with respect to the bias. Similarly, when using the GLASSO we see a consistent improvement with respect to the bias. For both cases, the GLASSO estimate improves over its sample covariance counterpart, even for higher $T$s.

The right-hand panel plots each estimator's variance, indicating the precision of the estimation. The non-regularized version not only faces heavy inaccuracies from the bias but also has an extremely elevated variance for $T < 150$. Its variance lessens with increasing $T$ but still underperforms compared to the regularized versions. For all regularization methods, we see gains in the variance, most likely stemming from the "bias-variance" trade-off. The combination of coefficient and covariance regularization therefore dominates for all $T$s. Pairing this finding with the findings of the left-hand panel, there appears to be no trade-off for $T < 200$ but an overall improvement in bias and variance. Overall, it is advisable to combine the regularizations for the coefficient and the covariance matrices as this combination provides the lowest variance and, for most sample sizes, it has also shown to come with the lowest bias.

Next, in the context of networks, we explicitly care about the diagnostic ability of the estimator. That is, we are interested in how well the classification into zero and nonzero entries performs in the network matrix. For that purpose, we summarize the performance in terms of the correct detection of zero and nonzero entries in the three panels in Figure 3. A value is considered true positive (TP) in the case of a hit and false positive (FP) in the case of a false alarm (Type I error). Likewise, a true negative (TN) is given if the FEVD is correctly estimated to be sparse at a given edge, and there is a false-negative (FN) when an existing edge is not found (Type II error). First, the probability of correct classification is summarized by the accuracy metric. Accuracy is defined as the number of true hits $(TP + TN)$ divided by the number of real positives and real negatives $N^2$, resulting in $accuracy = (TP + TN)/N^2$. We compare this metric of the estimators for increasing $T$. Clearly, a big improvement is unlocked when using the adaptive elastic-net estimator instead of OLS in small samples. A small but additional gain can be achieved through the usage of GLASSO instead of the covariance matrix. The differences largely disappear when $T$ increases.

The receiver operating characteristic (ROC) plots the false-positive rate $FPR = FP/(FP + TN)$ against the true-positive rate $TPR = TP/(TP + FN)$ while varying the discrimination threshold of setting values to zero. More precisely, the threshold varies from the lowest to the highest entry in the FEVD and, thus, sets increasingly more values to zero. For each threshold, the ROC plots the respective FPR and TPR in a diagram ranging from 0 to 1. A perfect estimator — i.e., one that correctly classifies all edges no matter the threshold — would result in a line starting at (x,y)=(0,1) and ending at (x,y)=(1,1). Conversely, a completely random guess would be shown as a 45° diagonal line. We plot the ROC curve for $T = 200$ in Panel B of Figure 3. Again, the regularized estimators clearly improve the performance when it comes to classification.

Finally, the area under the ROC curve summarizes the ROC curves in a single number in the right-hand panel, with higher values corresponding to better edge detection capabilities. The values range from 0.5 to 1, where 0.5 is a completely random guess and 1 is the perfect classification. This metric has the advantage of being able to summarize an estimate in a single number and, thereby, allows us to compare estimators for different $T$s. For $T = 100$, the regularized estimators already display high confidence in their classification performance. While the gains over the unregularized estimation mainly originate from the regression stage of the estimation, the results improve even further after regularizing the covariance matrix.

To sum up, our simulations demonstrate that the regularized estimators perform significantly better than their unregularized counterparts. This is evident when comparing estimation errors as well as when using entry-wise classification metrics. As expected, performance gains are most nuanced for small $T$s relative to $N$. However, our results indicate that the use of regularization also leads to improvements when estimating large-

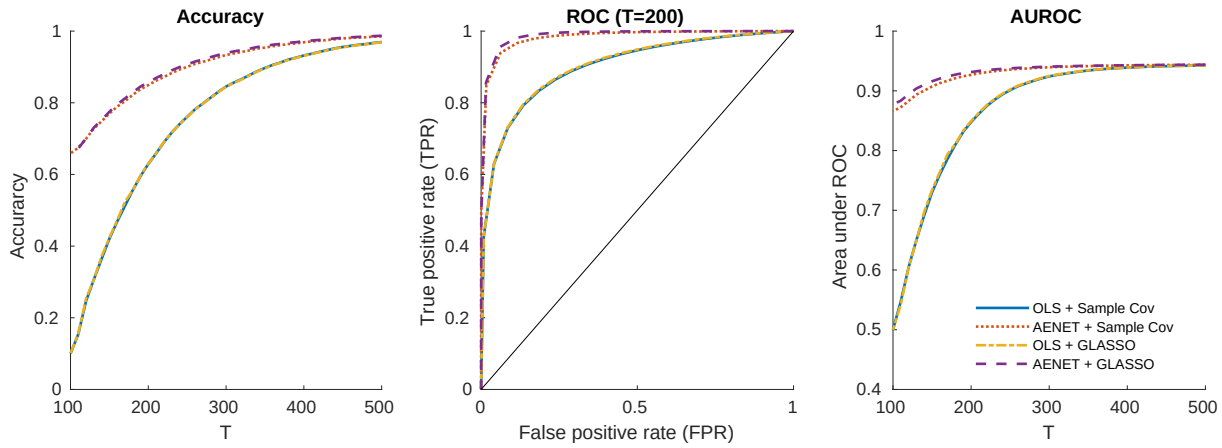


Figure 3: Simulation results for 500 Monte Carlo repetitions of *DGP 5* and $N = 100$. The first panel shows the accuracy for increasing sample size $T$ on the x-axis. The center panel shows the receiver operating characteristic (ROC) for $T = 200$, which has FPR on the x-axis and TPR on the y-axis. The diagonal thin black line is the equivalent of a random estimate. The right-hand panel shows the area under the ROC for increasing sample size $T$. AENET denotes the adaptive elastic-net.

dimensional networks from long datasets and, therefore, we advise researchers to apply regularization techniques when faced with large-dimensional estimation problems.

# 4 Empirical Application: Production Volatility Spillovers and the Great Moderation

The period known as the Great Moderation, starting in the mid-1980s, is characterized as a period with reduced fluctuations in many macroeconomic time series, such as real growth rates, industrial production (IP), and unemployment. For instance, the Federal Reserve Board's Index of Industrial Production shows a significant decrease in volatility after 1984 (see Figure 4). The question whether the Great Moderation happened due to good luck, better monetary policy, or any other structural change is vital for policy makers as they need to understand the impact of these factors on macroeconomic volatility.[12] Foerster et al. (2011) investigate this shift by decomposing IP into sectoral and common shocks. With the aggregated IP index summing over many weighted sector-level shocks, its large volatility is puzzling. The authors search for the roots of this puzzle and conclude that a decline in the volatility of common shocks induced most of the break in IP's volatility.

We are interested in understanding the source of the decline in the volatility of aggregate IP. Gabaix (2011) proposes that a small number of constituents with large index weights can explain aggregate shocks, an idea that Foerster et al. (2011) refutes. In our view, to analyze the influence of a particular industry on the aggregate, one needs to

[12]For an early discussion of the potential drivers of the Great Moderation see Bernanke (2004).

consider not only sectoral weights but also directional dependencies. In particular, we want to know whether a handful of large sectors lowered their links to other sectors and thus reduced common shock volatility and pairwise correlation. Acemoglu et al. (2012) establish this possibility based on input-output linkages and show that the diversification argument does not apply in the presence of strong network structures. Carvalho (2014) empirically investigates some implications of the production network in more detail and finds that pair-wise correlations between two sectors are bigger for stronger input-output links.

For insights into sectoral inter-connections, we obtain estimates of the directional dependencies. In contrast to a fast-expanding literature on production networks that exploits input-output tables as network proxies (for a review see Carvalho and Tahbaz-Salehi (2019)), our methodology directly sheds light on the network implied by sectoral IP correlations. An advantage of this approach is the ability to capture not only the supply-demand relationships of economic sectors but also alternative channels of transmission. For example, production volatility can also propagate between two sectors without an input-output relation if they compete for the same resources, if their outputs are substitutes or complements, or if technological innovations are transferable. Thus, by estimating FEVD tables, our application adds to the understanding of sectoral spillovers.

We examine pre-and post-1984 periods in detail to give a comparison of how spillovers have changed. Our findings provide novel empirical evidence in relation to the arguments of Foerster et al. (2011), Gabaix (2011) and Acemoglu et al. (2012) and offer a unifying view on their otherwise opposing narratives. We conjecture that sizeable spillovers from a handful of sectors initially increased the variance of aggregate shocks. With the Great Moderation, the structure of the spillover network changed such that sectoral shocks are less likely to amplify into aggregate volatility. Hence, our findings raise an alternative explanation where structural change largely contributed to the Great Moderation.

## 4.1 Data

As in Foerster et al. (2011), we use sectoral data on IP throughout the period 1972-2008. This data spans up to $N = 138$ sectors, which corresponds to the six-digit classification of the North American Industry Classification System (NAICS). The sectoral indices are available on a monthly basis. Since the pre- and post-Great Moderation periods have different sample sizes, they face distinct degrees of estimation uncertainty. Hence, we split the whole sample into three equally sized subsamples of $T = 144$ months. The subsamples span from 03/1972 to 02/1984, from 03/1984 to 02/1996, and from 03/1996 to 02/2008. In our analysis, the boundary between the first and second samples marks the onset of the Great Moderation. For simplicity, we label these samples as 1972-1983, 1984-1995, and 1996-2007, respectively.

Let $IP_{i,t}$ denote the value of IP of sector $i$ at date $t$. We take monthly growth rates and annualize the respective percentage points, $g_{i,t} = 1200 \times \ln(IP_{i,t}/IP_{i,t-1})$. The aggregate level of IP growth is the weighted average over the sectors, $g_t = \sum_{i=1}^{N} w_{i,t} g_{i,t}$, with given weights $w_{i,t}$. Figure 4 plots the growth rate of IP on an aggregate level. The first subsample, from 1972 to 1983, coincides with the pre-Great Moderation period and the other two subsamples are post-Great Moderation. It is evident that the average monthly volatility of aggregate IP diminished with the Great Moderation and stayed fairly constant thereafter.

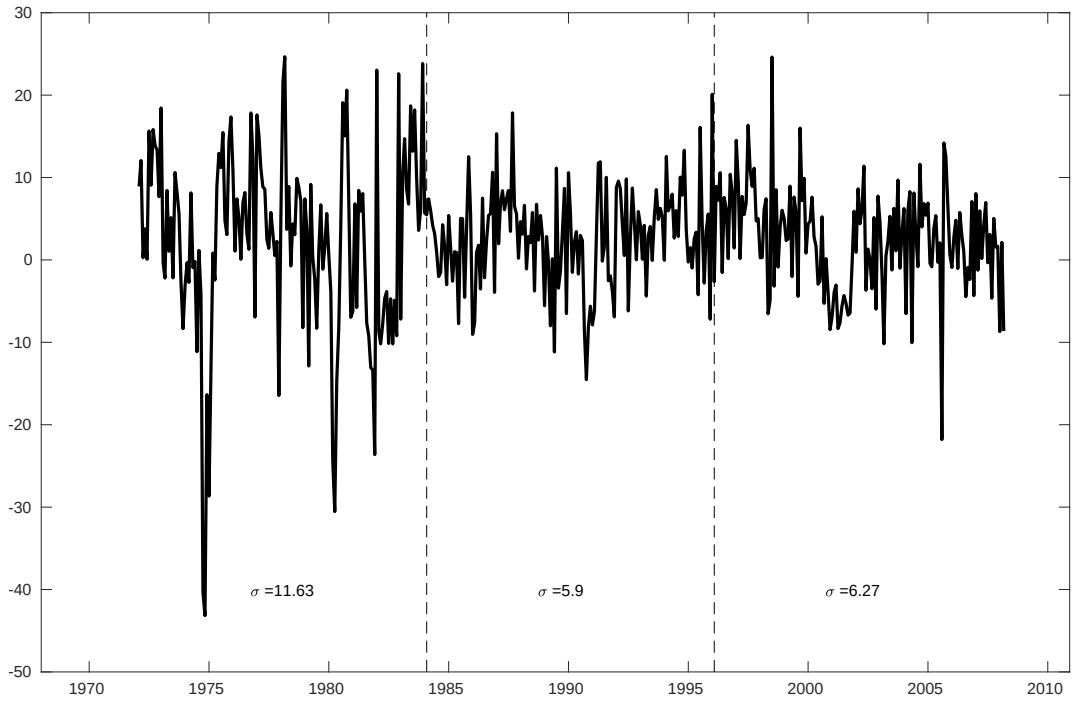


Figure 4: Annualized growth rates of monthly aggregate industrial production (IP) in percentage points. $\sigma$ denotes the sample volatility in the respective subperiod.

## 4.2 Estimation

We interpret the spillover constituent in the data as a VAR(1) model. We infer the connectedness in our model from the auto-correlation dynamics of $g_{i,t}$ in the monthly data. The higher frequency allows setting the forecast horizon to three months, which corresponds to the (undirected) covariance matrix of the quarterly data. This connection between monthly and quarterly frequencies gives insights into the contagion within a quarter. In particular, Foerster et al.'s (2011) average pairwise correlations and aggregate shocks may be better understood if we break up quarterly volatilities into three serially

correlated monthly volatilities. Hence, the central regression specification is as follows:

$$\begin{aligned} y_t &= \mu + Ay_{t-1} + u_t, \quad \forall t = 1, ..., T, \\ y_t &= [g_{1,t}, ..., g_{N,t}]', \\ u_t &\sim \mathcal{N}(0, \Sigma). \end{aligned}$$

We are aware of the shortcomings of this plain econometric specification. The frameworks proposed in Foerster et al. (2011) are tailored to test hypotheses about the common factors. However, our framework aims to see correlations through the lens of inter-sectoral spillovers as opposed to understanding correlations as common factors. While we agree with the original authors' approach to common variation, we deliberately set up our specification in sharp contrast in order to augment previous narratives, with the ultimate goal of providing a better understanding of the Great Moderation.

We regularize $A$ and $\Sigma$ with the techniques mentioned in the previous sections. Since the simulations did not point towards a consistent winner throughout all settings, we validate all regularization methods on the data. Namely, we run a 12-fold CV to select the best-performing estimator.[13] This step also includes the selection of $\alpha$ and $\lambda$ in the adaptive elastic-net and the hyperparameters $\lambda$, $\delta$ and $\eta$ for the covariance estimation. Then, we calculate the FEVDs with the forecast horizon $H = 3$. Recall that we consider the generalized version of FEVDs, i.e., the shocks are technically not idiosyncratic and can be correlated. Innovations that propagate via the auto-regressive matrix $A$ not only increase correlations between the sectors but also show directional effects that define the asymmetry of the FEVD.

We row-normalize $D^{gH}$ in (8) to show the percentage contribution to the variance. The row-normalized matrix is denoted as $\tilde{D}^{gH}$ with entries $\tilde{d}_{ij}^{gH}$. Additionally, we present key figures related to the network literature. In particular, we use the same measures as Diebold and Yılmaz (2014): in-, out-, and average connectedness. These measures are defined as the row sum, column sum, and the average row sum without the diagonal entries, respectively. To facilitate intuition, we slightly deviate from the original authors' terminology and henceforth use the terms *in-* and *out-connectedness* for what they call from- and to-connectedness.

$$C_{i\leftarrow\bullet}\left(\tilde{D}^{gH}\right) = \sum_{j\neq i} \tilde{d}_{ij}^{gH}, \qquad \text{(in-connectedness to i)}$$

$$C_{\bullet\leftarrow j}\left(\tilde{D}^{gH}\right) = \sum_{i\neq j} \tilde{d}_{ij}^{gH}, \qquad \text{(out-connectedness from j)}$$

$$C\left(\tilde{D}^{gH}\right) = \frac{1}{N}\sum_{i}\sum_{j\neq i} \tilde{d}_{ij}^{gH}. \qquad \text{(average connectedness)}$$

[13] In contrast to the statistical conventions, we use 12 folds to eliminate potential seasonalities.

The first two measures are sector-specific measures. The latter sums up the overall explanatory power of connectedness, which gives us an idea of how much variation is explained, on average, by spillovers and not directly by shocks. Note that we are mainly interested in the distribution of the outgoing spillovers of the sectors measured by the out-connectedness. Precisely, if a handful of large sectors had a high out-connectedness, the volatility of those sectors' IPs would not average out in the aggregate IP index.

First, we validate the performance of each method on the dataset. We use the three-digit sectoral disaggregation with 88 sectors. Figures 5 and 6 show the MSEs for the regularization of $A$ and $\Sigma$, respectively. As in the preceding simulations, we compare different methods for the estimation of the coefficient matrix $A$ and the covariance matrix $\Sigma$. For the estimation of $A$ we apply the adaptive elastic-net and compare the minimal MSE for different values of $\alpha$. Recall that $\alpha = 1$ is the adaptive LASSO and $\alpha = 0$ is the (adaptive) ridge estimator. $\lambda$ is chosen from 100 values spaced logarithmically between $10^{-5}$ and $10^{5}$. The grid for $\alpha$ includes $\{0, 10^{-5}, ..., 10^{-0.3}, 0.66, 0.75, 0.9, 1\}$ with 20 values spaced logarithmically between $10^{-5}$ and $10^{-0.3}$.

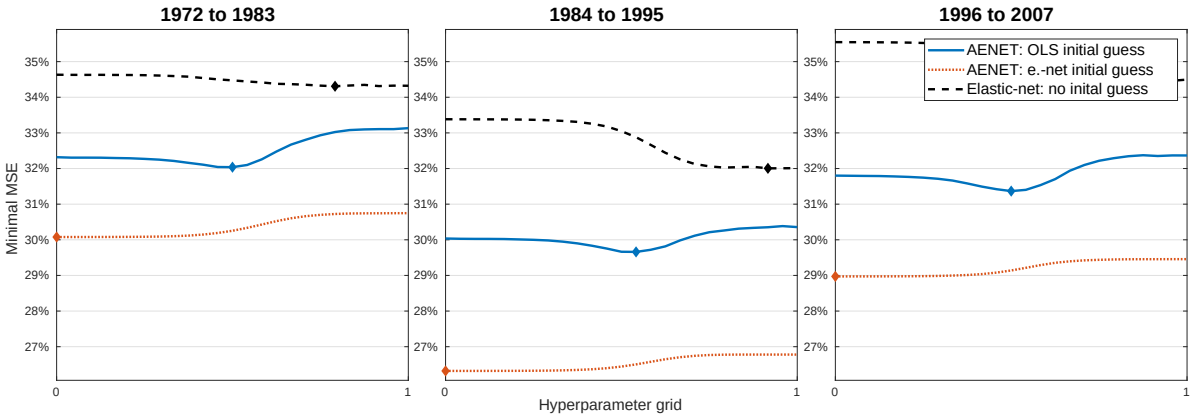


Figure 5: 12-fold CV results for the tuning parameter $\alpha$ in the adaptive elastic-net. The scale shows the percentage improvement over the OLS case, e.g., 33% refers to a forecast error being only a third that of the OLS estimate. The blue solid curves show the minimal MSE with $\hat{A}_{OLS}$ as the initial estimate in the weights of the individual penalties for different $\alpha$s, following the approach of Demirer et al. (2017). The red dotted curves show the same for $\hat{A}_{ENET}$ as an initial guess (selected by 12-fold CV, $\alpha = 0.5$, and $w_i = 1$), the black dashed lines correspond to the original (non-adaptive) elastic-net as in Zou and Hastie (2005).

Figure 5 shows the values of $\alpha$ on the x-axis and the minimized MSE (with respect to $\lambda$) on the y-axis. The dashed black lines represent simple elastic-net estimators, while the solid blue and dotted red lines are adaptive estimators with different initial guesses for the weight $w_i = |\hat{\beta}_{i,ini}|^{-1}$. Two results are evident here. First, the adaptive elastic-net estimator with elastic-net initialisation consistently offers an improvement over both the OLS-initialized adaptive elastic-net estimator and the plain elastic-net estimator for all values of $\alpha \in (0, 1)$. Second, as argued in Zou and Zhang (2009), the initialization of the adaptive elastic-net with plain elastic-net estimates already performs variable selection, such that the small alphas that focus on shrinkage are optimal with

this estimator. Consequently, CV suggests an estimator with a penalty close to ridge, and with the elastic-net estimate as an initial guess, resulting in $\alpha = 0$ for all periods.

In the second step, the CV for the covariance estimators includes the adaptive-thresholding estimator, the GLASSO estimator, the bona-fide Ledoit-Wolf estimator, and the Ledoit-Wolf estimator with a manually selected shrinkage weight. The latter is a heuristic extension of the Ledoit-Wolf estimator with manual selections of $b^2$ and $d^2$. Figure 6 shows their performance with different penalization and shrinkage parameters on the abscissa.

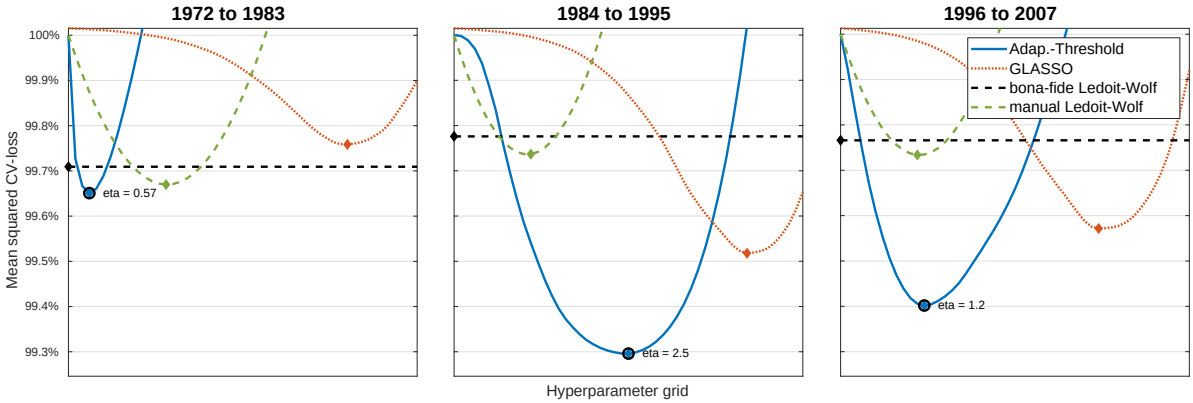


Figure 6: The 12-fold CV results for the tuning parameter of different covariance regularization methods. Values of the tuning parameters are on the x-axis and the MSEs are shown on the y-axis. The two thresholding estimators tune $\delta$ in (15), the GLASSO tunes $\lambda$ in (18) and the manual Ledoit-Wolf tunes the otherwise automatic shrinkage parameter $\delta$. The x-axis is normalized for each regularization method such that the convexity of all of the MSE curves is visible.

CV selects the adaptive-threshold estimator for all periods. This estimation procedure thresholds the values entrywise and appears to slightly outperform the others in the first period and distinctively outperform them in the two subsequent periods. The manual Ledoit-Wolf estimator comes second in the first period but falls back behind the GLASSO estimator later on. Our results from the simulation section provide a potential explanation for this ranking. With the respective estimators performing well for different DGPs in our simulation study, we see this shift in ranking as a sign that the pre-Great Moderation period may have a denser covariance matrix, which is in line with the findings of Foerster et al. (2011). In particular, the GLASSO approach sets the partial correlations between the two series to zero, which might be too restrictive for this subsample.

## 4.3 Results

Table 3 shows some selected statistics of the different estimations. Two results are evident in this table. First, the non-regularized version has a significantly higher average connectedness and stays remarkably constant over the periods. That is, it does not detect any change with the onset of the Great Moderation. In contrast, all of the regularized

| | | 1972-1983 | | 1984-1995 | | 1996-2007 | |
|---|---|---|---|---|---|---|---|
| | | $C(D^{gH})$ | *df* | $C(D^{gH})$ | *df* | $C(D^{gH})$ | *df* |
| Three-digit | non-regularized | 83.3% | | 82.3% | | 81.2% | |
| (88 sectors) | regularized (CV) | 47.6% | 10.3% | 27.9% | 10.4% | 34.1% | 10.7% |
| Four-digit | non-regularized | 92.6% | | 92.4% | | 92.0% | |
| (117 sectors) | regularized (CV) | 44.2% | 7.8% | 34.2% | 5.0% | 26.7% | 6.3% |
| Six-digit | non-regularized | 98.5% | | 98.5% | | 98.3% | |
| (138 sectors) | regularized (CV) | 42.9% | 6.4% | 27.5% | 4.7% | 30.3% | 5.6% |

Table 3: Summary of the estimation results for different levels of sectoral disaggregation. The columns labeled $C(D^{gH})$ show the estimated average connectedness. The columns labeled *df* show the density level of the autoregression coefficient matrix $A$, corresponding to the used degrees of freedom as a percentage of $N^2$.

versions capture the difference between the pre- and post-Great Moderation periods. Second, the average results of the regularized versions are robust over different levels of disaggregation. The non-regularized estimators, however, have higher average connectedness when we increase the dimensions. This observation clearly emphasizes the need of regularization in this context and exemplifies the bias that can occur in FEVDs (see Section 3).

Eventually, we want to emphasize the decreased usage of the degrees of freedom of the regularized estimators in Table 3. For robustness, we also report the results for lower levels of sectoral aggregation. As expected, the fractional degrees of freedom in the regularized coefficient estimates $\hat{A}_{reg}$ decrease with higher dimensions since the estimation has to deal with more estimation uncertainty and larger networks are likely more sparse. However, with the estimated coefficient matrices being reasonably populated and the network connectedness staying fairly unchanged across the specifications, we conclude that the three-digit aggregation level offers a reasonable balance between granular insights and keeping high dimensionality at bay. In the analysis, we observe average spillovers of 47.6%, 27.9%, and 34.1% for the three periods, respectively.

To give a holistic view of the network, we summarize the estimated row-normalized connectedness tables in Figure 7. The figure shows the tables as network graphs, in which the force-directed graph drawing algorithm arranges the nodes. That is, two nodes appear closer in the graph if they have stronger connections to each other. The size of the node relates to the respective average weight $\bar{w}_{i,t^*}$ of the sector in the IP index. The colors illustrate the out-connectedness. Finally, we label the sector with the highest out-connectedness.

From eye-balling, it is evident that the network significantly changed after 1984. Whereas the pre-Great Moderation period shows a closely connected graph with a handful of powerful nodes in the center, the two consecutive periods have more widespread graphs with less concentration. Clearly, this also shows in the average connectedness

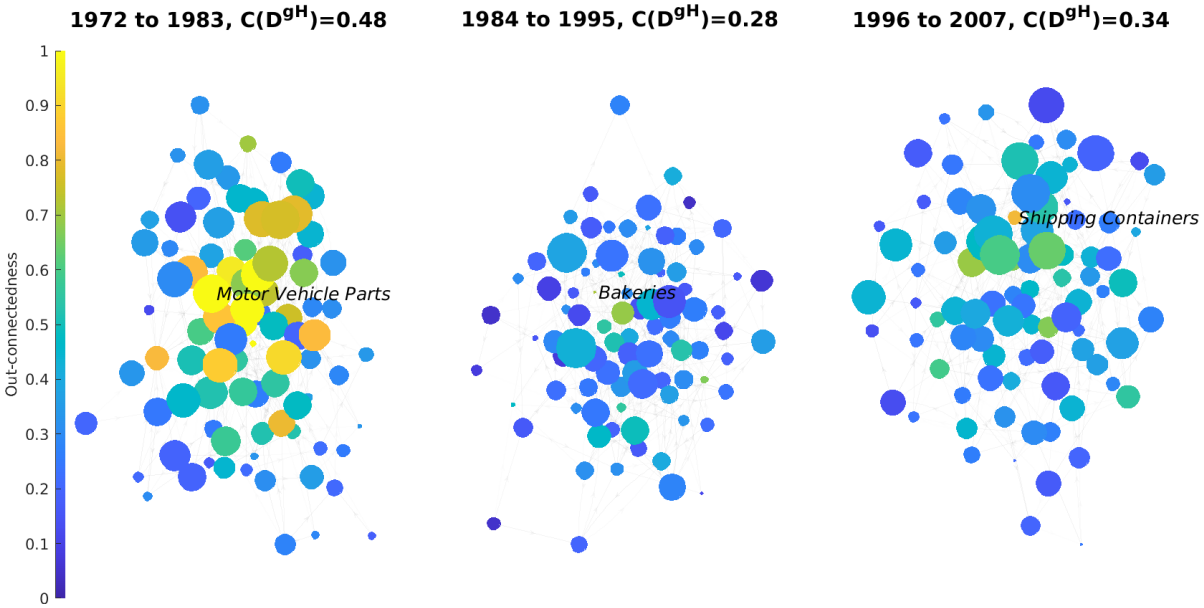


Figure 7: Connectedness networks for the respective periods. The size of the node relates to the respective weight of the sector in the IP index. The colors depict the out-connectedness. For the sake of the visualization, we cap the color scale at 1. The sectors with the highest out-connectedness are labeled.

$C(D^{gH})$. This result alone can easily be consolidated with the explanations in Foerster et al. (2011). Adding to this, we now investigate the asymmetric part of the spillover tables by looking at sectoral in- and out-connectedness. As mentioned earlier, a plausible explanation for the high volatility in the aggregate index is that a handful of sectors spilled a lot of volatility before the Great Moderation and decreased contributions afterwards. As a first observation from Figure 7, large values for out-connectedness became less frequent and sectors with smaller weights (smaller nodes) took more-central positions in the networks after 1984.

Although more spillovers do not necessarily result in higher aggregate volatility, the root may be in the concentration of outgoing spillovers. In that regard, out-connectedness measures how much a single sector's volatility explains the volatility of all other sectors. If a handful of sectors have very high levels of out-connectedness, then their volatility determines, to a large extent, the volatility of other sectors. Consequently, the volatility of the aggregate IP index is also indirectly affected by innovations to those sectors and their shocks do not average out.

Figure 8 displays the counter-cumulative density functions (CCDFs) of the out-connectedness measure for the three sample periods. Notably, the distribution in the pre-Great Moderation period from 1972 to 1983 expands much further to the right than in subsequent periods. The shift of the curve to the right alone can be explained by the larger correlations. However, the pronounced right tail for this period is an indication of a heightened network concentration, supporting the hypothesis of network-induced under-diversification of the aggregate IP index. Roughly 10% of the sectors have an out-connectedness higher than 1, meaning that each explains more than 100% of the volatility

of the other sectors in total. Therefore, their shock volatilities would show up more than twofold in an aggregate index with equal weights. The fat tail mostly disappears after the Great Moderation, i.e., the subsequent periods have a more-diversified production network. Lastly, we observe that the left tail of the CCDF also experienced a shift of the out-connectedness with a smaller magnitude, which reflects a reduction in connectedness overall.[14]

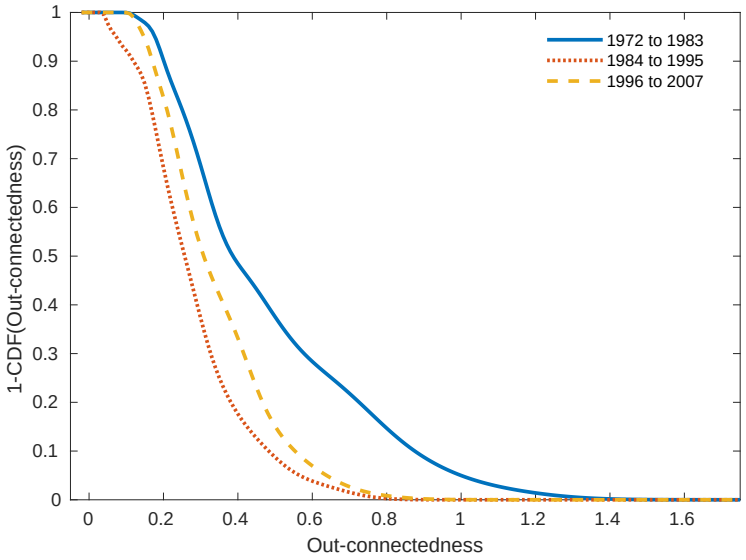


Figure 8: Estimated counter-cumulative density functions (CCDFs) of the out-connectedness of the 88 three-digit-level sectors.

Figure 9 tracks the dynamics of the out-connectedness for individual sectors. Each dot in the plot corresponds to a sector, with the position along the y- and x-axes according to their out-connectedness before and after the Great Moderation, respectively. The size of the dot relates to the size of the sector in the IP index. The colors relate to the changes in each sector's out-connectedness. Sectors above the 45° line have decreased their effect on other sectors and vice versa. The shift in distribution is immediately evident as most sectors appear in the top left of the graph. However, this graph clearly shows a strong decrease in a couple of sectors (colored in dark blue), whereas the out-connectedness stayed relatively constant for other sectors (red and yellow). Remarkably, it also seems that some sectors increased in out-connectedness. This finding supports the hypothesis of structural change as a contributor to the Great Moderation.

In addition to sectoral out-connectedness, the distribution of in-connectedness may have played a critical role in the structural change that emerged during the Great Mod-

[14]Supporting histograms of the in- and out-connectedness measures (Figures D.13 and D.14) are available in Appendix D. Here, the 10% strongest sectors by pre-Great Moderation out-connectedness have mostly reduced their values to around 0.5 in the last period. Comparing only the first and last periods, the distributions are similar except for the heavy right tail.

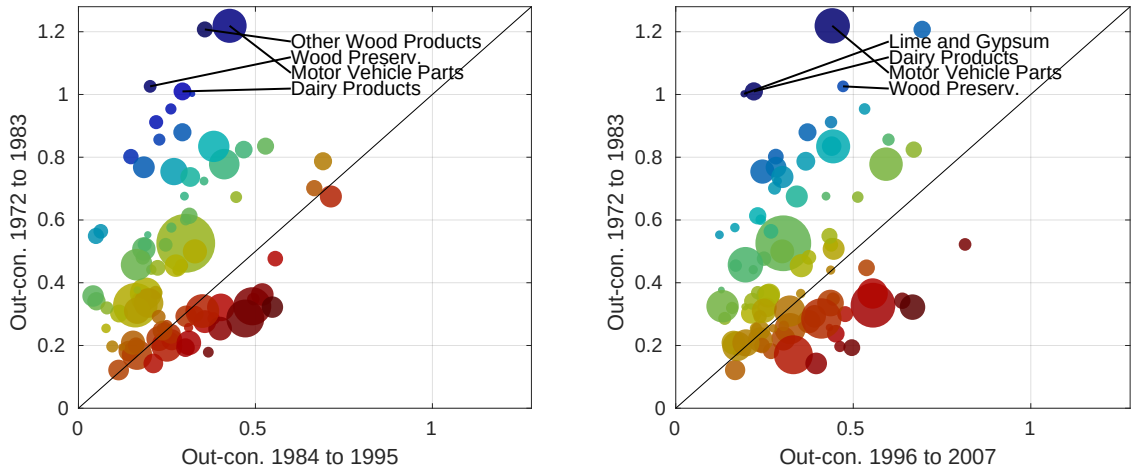


Figure 9: Out-connectedness scatter plot of the 88 three-digit-level sectors pre- versus post-Great Moderation. Values on the $y$- and $x$-axes show the out-connectedness for pre- and post-Great Moderation periods, respectively. The size of the markers correspond to the size of the sectors. The colors show how far away the sectors are from the diagonal line, i.e., how much their out-connectedness changed relative to the pre-Great Moderation period. The four sectors with the largest decreases in out-connectedness are labeled.

eration. However, we find its role in amplifying volatility to be ambiguous because of two channels. First, big values of in-connectedness imply higher dependencies of other sectors such that a small sector's volatility may additionally emerge in sectors with bigger weights. That is, if we see a handful of sectors with large weights having very strong in-connectedness, the spillovers might skew the weighting scheme of the aggregate index. Second, higher values of in-connectedness are generally related to a higher degree of diversification, which in return would wash out the intrinsic volatility contribution. Paired with an infelicitous distribution of weights — e.g., sectors with big weights might have received more spillovers after 1984 — this could lead to more diversification overall. This latter hypothesis is in line with the argument of Gabaix (2011) that a small number of constituents with large index weights can explain aggregate shocks. We analyze both conjectures in Appendix D in Figure D.12 but do not find support for any of the aforementioned.

Next, we want to understand how the full network structure affected aggregate index volatility and how individual sectors contributed to the aggregate. Take the aggregate index representation from before: $g_t = \sum_{i=1}^{N} w_{i,t} g_{i,t}$; the IP index variance is $\sigma_p^2 = \bar{w}' \Sigma \bar{w}$, where $\bar{w}$ is a vector of average weights in the respective period and $\Sigma$ is the covariance matrix of the index constituents. Now, we leverage the higher frequency of our estimation, which ultimately allows us to have directed estimates of variance spillovers. The sum of the monthly inter-temporal spillovers then determines the lower-frequency estimate. Thus, we use the sum of the three-month covariance of observations to innovations $IRV = \sum_{h=0}^{2} \Phi_h \Sigma = \sum_{h=0}^{2} cov(y_t, u_{t-h})$, where $\Phi_h = A^h$ as in (2). We denote this term as the innovation response variance (IRV) and define the vector $v_{\bullet \to IP}$ as the decomposition of

the aggregate IP index variance into sectoral contributions:

$$v_{\bullet\to IP} = \bar{w}' IRV diag(\bar{w}), \quad \text{and } v^{agg} := \mathbf{1}' v_{\bullet\to IP}, \tag{19}$$

This representation is particularly appealing, as, for example, entry $i, j$ in the IRV matrix shows how much variation in variable $i$ follows from variations of innovations to $j$. Note that in our framework, the IRV is a full description of the variance in $y_t$ as $u_{t-h}$ is the only source of variation and comprises both sector-level and aggregate innovations. Thus, $v_{\bullet\to IP}$ accounts for the variance, the weight of the sector, and the magnitude of the spillovers to other sectors. Note that the vector $v_{\bullet\to IP}$ shows the dissection of the quarterly data by allowing directional reactions to occur at monthly frequencies. Further, we decompose this representation in

$$v_{\bullet\to IP} = \underbrace{\bar{w}' diag\left(IRV\right) diag(\bar{w})}_{\text{intrinsic component}} + \underbrace{\bar{w}' \left[IRV - diag\left(IRV\right)\right] diag(\bar{w})}_{\text{extrinsic component}}. \tag{20}$$

This decomposition into intrinsic and extrinsic components allows us to understand the channels through which the sectors contributed to the variance of the index, thus, providing additional insight into the drivers of the Great Moderation, as discussed in the following.

Table 4 contains the decomposition of the aggregate variance into multiple components. First, we confirm that the aggregate variance, $v^{agg}$, implied by our estimates matches the magnitude of the data variance, $v^{IP}$. Furthermore, the decomposition of the aggregate variance into an intrinsic component, $v^{IC}$, and an extrinsic component, $v^{EC}$, shows a decrease in both components with the onset of the Great Moderation. Interestingly, we see a sharper drop in the extrinsic component from 61 to 3.9. This finding is generally in line with the findings of Foerster et al. (2011). An additional decomposition of the extrinsic component into a symmetric, $v^{Sy}$, and an asymmetric portion, $v^{Asy}$, allows us to shed light on the credibility of the network story as opposed to the narrative of aggregate shocks. Although the symmetric portion seems to be the dominant force of the extrinsic component, we still see a non-negligible decrease in the asymmetric portion. Perhaps surprisingly, the asymmetric shares, $v^{Asy}/v^{agg}$ and $v^{Asy}/v^{EC}$, drop more considerably than the symmetric counterpart after 1984. We see this as support of the structural change in the network contributing to the Great Moderation.

Lastly, we care about single-sector contributions via the IRV in (19). To do so, Figure 10 tracks the sectors' variance contributions over the three periods. We place the sectors with the highest contributions at the bottom and vice versa. In brackets, we show the percentage of the sectors' contributions via the extrinsic channel as defined in (20). A higher value relates to more systemic relevance in the aggregate index.

First, we note that the sector labeled Motor Vehicles is estimated to be the biggest

| | | 1972-1983 | 1984-1995 | 1996-2007 |
|---|---|---|---|---|
| Aggregate variance | $v^{agg} = \bar{w}' IRV \bar{w}$ | 81.1 | 14.1 | 19.9 |
| Data variance | $v^{IP} = var(\bar{w}' y_t^q)$ | 94.6 | 14 | 14.6 |
| Intrinsic component | $v^{IC} = \bar{w}' diag(IRV) \bar{w}$ | 20.2 | 10.1 | 11.8 |
| Extrinsic component | $v^{EC} = v^{agg} - v^{IC}$ | 61 | 3.9 | 8.1 |
| of which symmetric | $v^{Sy} = \bar{w}'(IRV^s - diag(IRV^s))w$ | 44.4 | 3.5 | 5.4 |
| of which asymmetric | $v^{Asy} = v^{EC} - v^{Sy}$ | 16.5 | 0.5 | 2.7 |
| Extrinsic share | $v^{EC}/v^{agg}$ | 0.75 | 0.28 | 0.4 |
| Asymmetric share | $v^{Asy}/v^{agg}$ | 0.2 | 0.03 | 0.14 |
| Asymmetric/extrinsic | $v^{Asy}/v^{EC}$ | 0.27 | 0.12 | 0.34 |

Table 4: Decomposition of variance contributions by channel. Aggregate variance is an estimate of the variance of the aggregate IP index via impulse response variances. The intrinsic component consists of all sectors' direct contributions to the index variance. The extrinsic component comprises all effects on the index variance through the channel of inter-sector linkages. We then further decompose the extrinsic component into a symmetric channel, which is common in $IRV$ and $IRV'$. The exact formula used to calculate the symmetric portion of $IRV$ is $IRV^s = \text{sign}[\text{sign}(IRV) + \text{sign}(IRV')] \odot \min\{\text{abs}(IRV), \text{abs}(IRV')\}$, where $\odot$ denotes the Hadamard product. Thus, off-diagonal entries are set to zero if the directional effects differ in sign; they are set to the value closer to zero if the directional

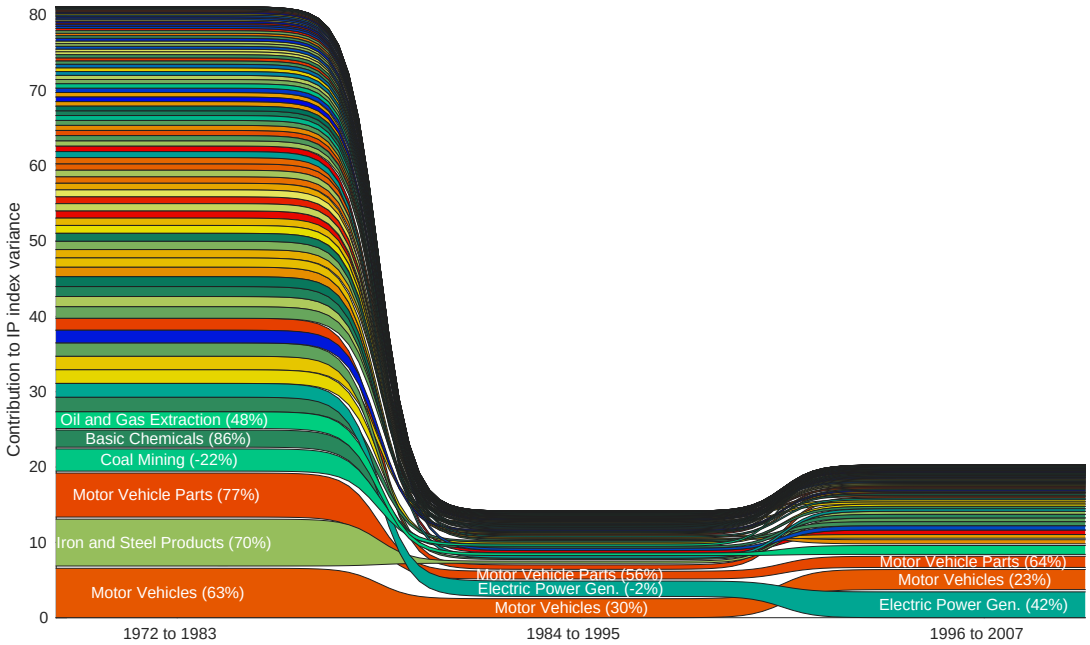


Figure 10: Bump chart of sectoral contributions to the variance of the aggregate IP index. Contributions are calculated as in (19). Data includes IP indices for 88 three-digit-level sectors. The top contributors per period are labeled. Values in brackets correspond to the extrinsic component of the contribution. For example, the value of the Motor Vehicles sector in the first period states that 63% of this sector's contribution is due to spillovers (extrinsic) to other sectors and 37% is from its intrinsic component.

contributor to the IP index variance before the Great Moderation. After 1984, we see nearly all contributions fade, with Motor Vehicles still having the biggest effect but a much lower extrinsic component. Perhaps surprisingly, Coal Mining is ranked fourth before the Great Moderation. With its variation primarily influenced by the sector labeled Oil and Gas Extraction (number not in the graph), the Coal Mining sector serves as an amplifier

of oil shocks. However, with an extrinsic component of only $-22\%$, the effect of the Coal Mining sector on the aggregate is partially diluted. A major change in sectoral contribution can be seen in Iron and Steel Products, a sector that almost exclusively contributes to the IP index through spillovers (70%), thereby increasing common variation across sectors. This sector's contribution dropped drastically with the Great Moderation. Similarly, Foerster et al. (2011) report that metal-related industries, in particular, are best explained by a factor structure, indicating that innovations in Iron and Steel Products could manifest in aggregate variation.

Throughout the application, some findings appeared puzzling at first; e.g., the large decrease in the effect of lumber-related sectors and Dairy Products to other sectors in Figure 9, and the strong but diluted effect of the Coal Mining sector on the aggregate index in Figure 10. In particular, the latter exhibited an unusually high variance before the Great Moderation. A possible hypothesis relating these three puzzles may be the prevalence of price and quantity controls in the period from 1972 to 1983. For example, in line with the comprehension of the Arrow-Debreu model, we conjecture that the suppressed IP variance in the Oil and Gas Extraction sector surfaced in other sectors, and particularly in Coal Mining. Hence, we believe that sizeable demand fluctuations for energy-related commodities could have had a great impact on the coal sector, which had a large effect on the aggregate index. In January 1981, an executive order mitigated this impact by re-allowing the market to freely adjust oil prices. In line with this explanation, from the second sample period onward, we additionally observe that sectors which previously faced price and quantity controls such as Dairy Products and Wood Preservation also lowered their out-connectedness in Figure 9. The coincidence of these developments warrants further research on the relationship between market interventions and the Great Moderation.

In summary, the results provide insights into the changes observable during the Great Moderation. Foerster et al. (2011) emphasize that the decrease in aggregate variance is not due to some sectors but instead is rooted in the change of aggregate shocks; i.e., shocks to multiple sectors at once. However, our analysis of the propagation of sector-specific shocks within a three-month period highlights that a handful of sectors may have largely contributed to the change in quarterly aggregate shocks. These sectors had strong spillover effects on other sectors such that they increased the variance of the aggregate index. In contrast to Gabaix (2011), not only is the index weight a pivotal characteristic for a sector to influence aggregate variance but also critical are the spillover effects it has on other constituents. Our results therefore offer a unifying perspective to the opposing explanations in Gabaix (2011) and Foerster et al. (2011), where the prevalence of aggregate shocks has fallen due to a decrease in decisive intersectoral spillovers.

# 5 Conclusions

In this paper, we investigate the estimation of high-dimensional vector autoregressive models. In a simulation study, we compare different regularization methods for the coefficient and the covariance matrix. We evaluate their performance in the estimation of forecast error variance decompositions and we find that the regularization of both matrices leads to better estimation. Since there is no single best estimator among all of the simulations, we suggest validating the estimators through cross-validation. In an application to US industrial production, we are able to uncover changes in the structure of the inter-sectoral spillover network around the Great Moderation. Specifically, we aim to answer the question whether a handful of sectors were responsible for the decrease in aggregate variance. We find that a couple of sectors had a particularly high outgoing spillover before the Great Moderation and their importance was unmatched thereafter.

# References


Acemoglu, D., Carvalho, V. M., Ozdaglar, A. and Tahbaz-Salehi, A. (2012), 'The network origins of aggregate fluctuations', *Econometrica* **80**(5), 1977–2016.

Barigozzi, M. and Brownlees, C. (2013), 'Nets: Network estimation for time series', *Journal of Applied Econometrics* .

Bergmeir, C., Hyndman, R. J. and Koo, B. (2018), 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', *Computational Statistics & Data Analysis* **120**, 70–83.

Bernanke, B. (2004), 'The great moderation', *Washington, DC* .

Bickel, P. J. and Levina, E. (2008), 'Regularized estimation of large covariance matrices', *The Annals of Statistics* pp. 199–227.

Cai, T. and Liu, W. (2011), 'Adaptive thresholding for sparse covariance matrix estimation', *Journal of the American Statistical Association* **106**(494), 672–684.

Carvalho, V. M. (2014), 'From micro to macro via production networks', *Journal of Economic Perspectives* **28**(4), 23–48.

Carvalho, V. M. and Tahbaz-Salehi, A. (2019), 'Production networks: A primer', *Annual Review of Economics* **11**, 635–663.

Demirer, M., Diebold, F. X., Liu, L. and Yilmaz, K. (2017), 'Estimating global bank network connectedness', *Journal of Applied Econometrics* .

Diebold, F. X. and Yılmaz, K. (2014), 'On the network topology of variance decompositions: Measuring the connectedness of financial firms', *Journal of Econometrics* **182**(1), 119–134.

Foerster, A. T., Sarte, P.-D. G. and Watson, M. W. (2011), 'Sectoral versus aggregate shocks: A structural factor analysis of industrial production', *Journal of Political Economy* **119**(1), 1–38.

Friedman, J., Hastie, T. and Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**(3), 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1.

Gabaix, X. (2011), 'The granular origins of aggregate fluctuations', *Econometrica* **79**(3), 733–772.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), 'The elements of statistical learning: data mining, inference, and prediction, springer series in statistics'.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

James, W. and Stein, C. (1992), Estimation with quadratic loss, *in* 'Breakthroughs in statistics', Springer, pp. 443–460.

Kappen, H. J. and Gómez, V. (2014), 'The variational garrote', *Machine Learning* **96**(3), 269–294.

Kascha, C. and Trenkler, C. (2015), Forecasting vars, model selection, and shrinkage, Technical report, Working Paper Series, Department of Economics, University of Mannheim.

Koop, G., Pesaran, M. H. and Potter, S. M. (1996), 'Impulse response analysis in nonlinear multivariate models', *Journal of Econometrics* **74**(1), 119–147.

Lam, C. et al. (2016), 'Nonparametric eigenvalue-regularized precision or covariance matrix estimator', *The Annals of Statistics* **44**(3), 928–953.

Ledoit, O. and Wolf, M. (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of Multivariate Analysis* **88**(2), 365–411.

Lütkepohl, H. (2005), *New introduction to multiple time series analysis*, Springer Science & Business Media.

Pesaran, H. H. and Shin, Y. (1998), 'Generalized impulse response analysis in linear multivariate models', *Economics letters* **58**(1), 17–29.

Qian, J., Hastie, T., Friedman, J., Tibshirani, R. and Simon, N. (2013), 'Glmnet for matlab', `http://www.stanford.edu/~hastie/glmnet_matlab/`.

Raveh, A. (1985), 'On the use of the inverse of the correlation matrix in multivariate data analysis', *The American Statistician* **39**(1), 39–42.

Rothman, A. J., Levina, E. and Zhu, J. (2009), 'Generalized thresholding of large covariance matrices', *Journal of the American Statistical Association* **104**(485), 177–186.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the royal statistical society: series B (statistical methodology)* **67**(2), 301–320.

Zou, H. and Zhang, H. H. (2009), 'On the adaptive elastic-net with a diverging number of parameters', *The Annals of Statistics* **37**(4), 1733.

# Online Appendix

## A Conditions for the thresholding rule

| | | |
|---|---|---|
| $(i)$ | $\lvert s_\lambda(z)\rvert \leq c\lvert y\rvert \quad \forall \quad z, y$ satisfying $\lvert z-y\rvert \leq \lambda$ and some $c>0$ | |
| $(i)^*$ | $\lvert s_\lambda(z)\rvert \leq \lvert z\rvert$ | (shrinkage) |
| $(ii)$ | $s_\lambda(z)=0$ for $\lvert z\rvert \leq \lambda$ | (thresholding) |
| $(iii)$ | $\lvert s_\lambda(z)-z\rvert \leq \lambda \quad \forall \quad z \in R$ | (limits of shrinkage) |

Cai and Liu (2011) conduct analyses for the class satisfying (i),(ii),(iii) but state that it is also possible to adapt this to (i)*. The hard-thresholding rule is ruled out by (i), but other thresholding functions, such as the soft-thresholding $s_\lambda(z) = \text{sgn}(z)(|z|-\lambda)_+$ and the adaptive LASSO rule $s_\lambda(z) = z(1-|\lambda/z|^\eta)_+$ with $\eta \geq 1$, are included.

## B Shrinking diagonal values in the Adaptive-Threshold Estimator

We experienced throughout the paper that some hyperparameters choose a shrinkage, which sets the diagonal values of the covariance matrix to zero. To counteract this phenomenon, we treat these values differently. Similar to the Ledoit-Wolf estimator, we do not shrink diagonal values $\hat{\sigma}_{ii}$ towards zero but towards the no-variance estimator $m = tr(\hat{\Sigma})N^{-1}$. More specifically, we shrink diagonal values with the following rule:

$$s_\lambda(\hat{\sigma}_{ii}) = \hat{\sigma}_{ii}(1-|\lambda/z|^\eta)_+ + m \cdot \max\{|\lambda/z|^\eta, 1\}, \qquad \forall i.$$

The term $\max\{|\lambda/z|^\eta, 1\}$ ranges between zero and one and has the opposite weight of $(1-|\lambda/z|^\eta)_+$. Thus, this rule chooses a linear combination from the no-bias full-variance estimator, $\hat{\sigma}_{ii}$, and the no-variance full-bias estimator, $m$. In the simulations, as well as in the CV, this treatment outperformed shrinkage towards zero.

# C Illustration of Cross-Validation Procedure

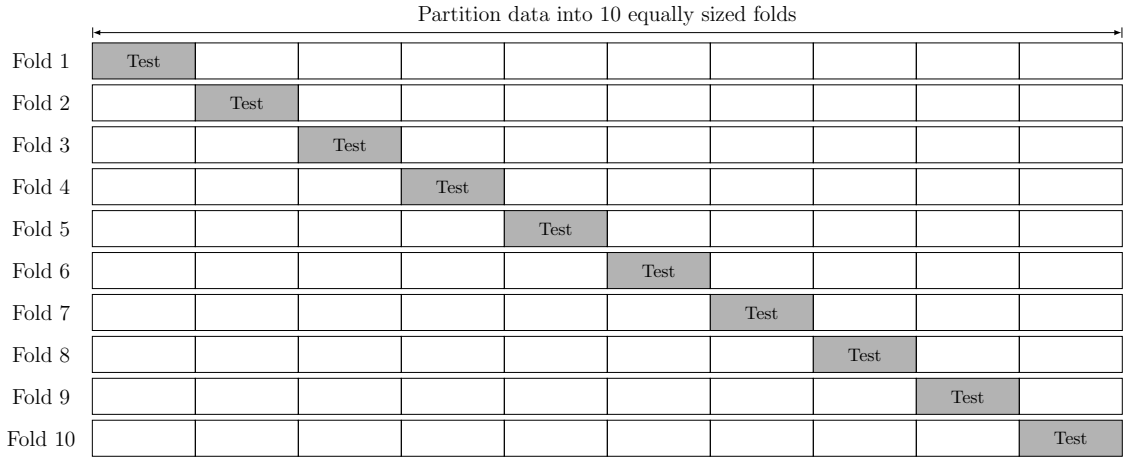


Figure C.11: Illustration of 10-fold CV. First we split the sample into 10 folds of equal length. Then we use 9 folds for training and validate the trained model's performance on the hold-out test fold.

# D Complementary Graphs

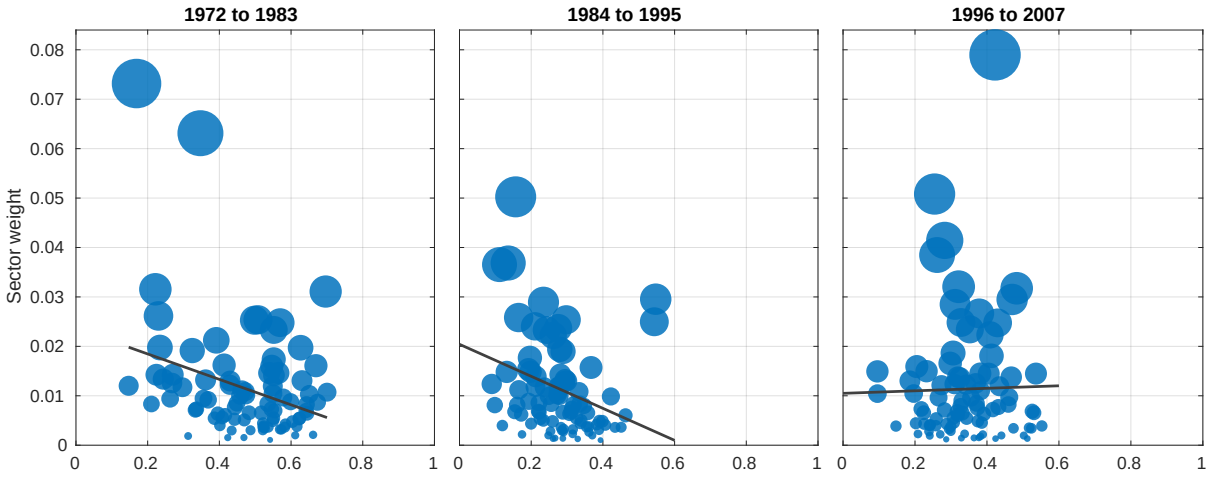


Figure D.12: Scatterplots of the in-connectedness of the 88 three-digit-level sectors against the sectoral weights in the aggregate IP index. The in-connectedness relates to incoming spillovers from other sectors. We plot the regression line per sample.

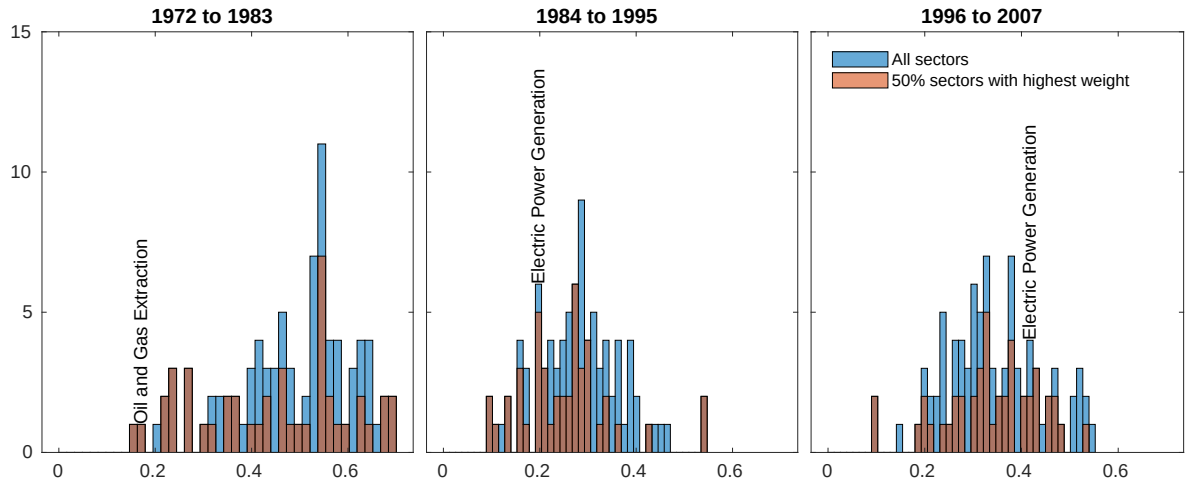


Figure D.13: Histogram of the in-connectedness of the 88 three-digit level sectors. The 50% of the sectors with the biggest weights are highlighted in red. The in-connectedness relates to incoming spillovers from other sectors. We follow the strongest sector by weight before the Great Moderation by labeling it throughout the samples.

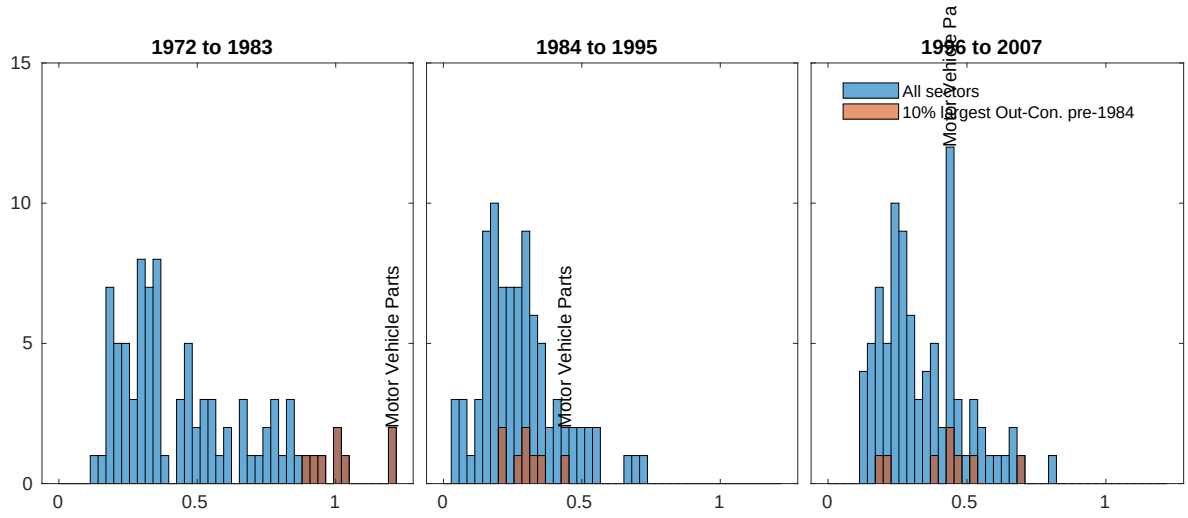


Figure D.14: Histogram of the out-connectedness of the 88 three-digit-level sectors. The 50% of the sectors with the biggest weights are highlighted in red. The out-connectedness relates to incoming spillovers from other sectors.

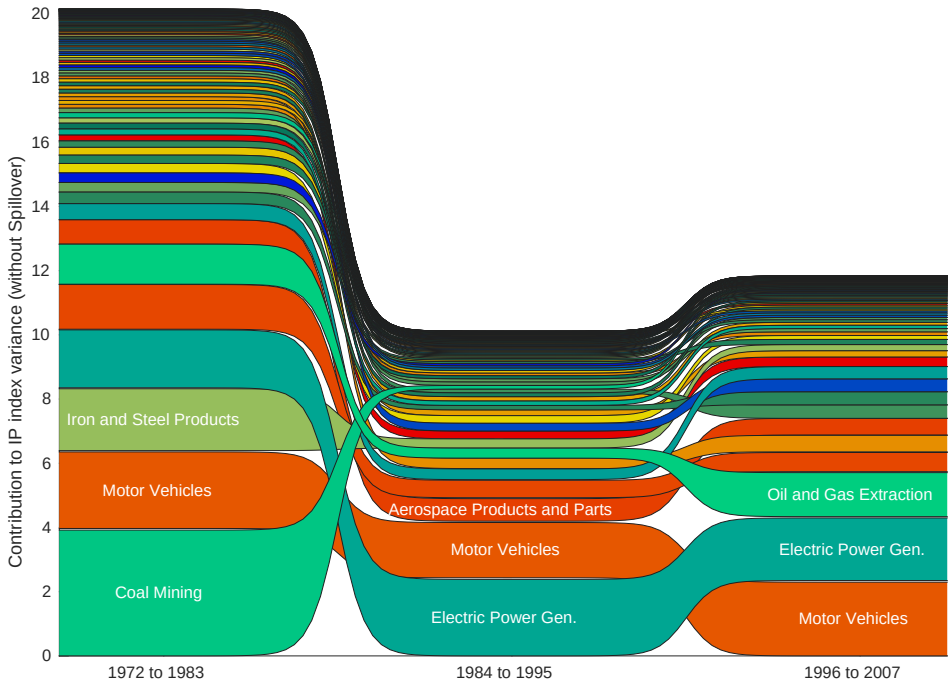


Figure D.15: Bump chart of sectoral contributions to the aggregate variance of the IP index without spillovers. Contributions are calculated as in (19). Data includes IP indices for 88 three-digit-level sectors. The top three contributors per period are labeled.

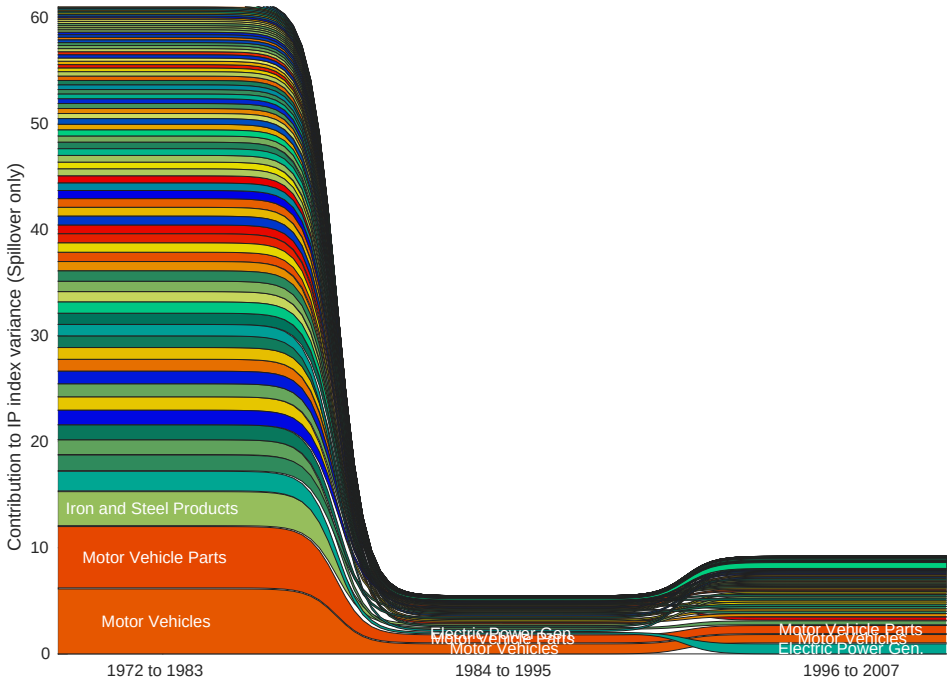


Figure D.16: Bump chart of sectoral contributions to the aggregate variance of the IP index with spillovers only. Contributions are calculated as in (19). Data includes IP indices for 88 three-digit-level sectors. The top three contributors per period are labeled.